IDENTIFICATION AND CHARACTERIZATION OF THE MAJOR GENE *MA* AND ITS ASSOCIATED CO-EXPRESSION GENE NETWORKS REGULATING APPLE FRUIT ACIDITY

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

In Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Yang Bai

August 2014

IDENTIFICATION AND CHARACTERIZATION OF THE MAJOR GENE *MA* AND ITS ASSOCIATED CO-EXPRESSION GENE NETWORKS REGULATING APPLE FRUIT ACIDITY

Yang Bai, Ph. D.

Cornell University 2014

Apple fruit acidity which considerably affects fruit taste and flavor is primarily determined by malate concentrations. Previous studies reported that apple fruit acidity was predominantly controlled by the major QTL *Ma*. To better understand apple fruit acidity, this study attempts to identify and characterize the gene underpinning *Ma* and its associated co-expression gene networks. To achieve the goal, three sets of experiments were conducted. The first set of experiments was designed to identify the candidate gene for *Ma* and was summarized in Chapter 1. This chapter presents that the *Ma* locus physically spans over 65kb and harbors two aluminum-activated malate transporter (ALMT) -like genes, designated *Ma1* and *Ma2*. Other important findings include 1) only *Ma1* was expressed in significant correlation with the variation of acid levels, and 2) there is a nucleotide mutation that leads to a pre-mature stop codon in *Ma1*.Overall, it concludes that *Ma1* rather than *Ma2* is the gene underlying *Ma*.

The second set of experiments aims at identification of the *Ma1* associated co-expression gene network regulating malate levels in developing fruit of Golden Delicious and was reported in Chapter 3. One finding of this chapter is that malate-pyruvate interconversion, photosynthesis, mitochondrial electron transport, and amino acid degradation were important pathways for the change of malate levels in developing fruit. The other finding is that symplastic signal transduction,

transcriptional regulation, post-translational modification and apoplastic roles were important in the regulation of fruit acidity.

The third set of experiments was conducted to identify the *Ma1* associated co-expression gene network controlling acidity in mature fruit of ten diverse apples and was described in Chapter 4. Data in this chapter suggested that calcium signaling was likely a crucial mechanism that regulates both the expression of *Ma1* and the *Ma1* associated co-expression gene network governing fruit acidity.

Chapter 2 is about improving the current version of apple reference transcriptome due to its unaccep low coverage in mapping of RNA-seq reads. The improved apple reference transcriptome comprises 71,178 genes (17,524 are novel) and increases the coverage of RNA-seq reads from 37-46% to 62-82%. This chapter lays a foundation for data analysis in Chapters 3 and 4. In conclusion, this study takes the understanding of apple fruit acidity to a higher level and opens more grounds for further dedicated studies.

BIOGRAPHICAL SKETCH

Yang Bai was born and brought up in Jinan, the center of Ji-Lu culture, on the northeast coast of China. When she was young, she was a quiet, tall, day-dreaming girl who had no other major habits but reading and painting. Her favorite books were the series of *Voyages Extraordinaires* by Jules Verne and the *Sherlock Holmes* novels by Sir A.C. Doyle, in which the beauty of nature, the smartness of innovation and the power of logical reasoning deeply attracted her and probably initiated her long-lasting interest in scientific research. She attended Shandong Normal University and graduated in 2007 with a B.S. in Biotechnology. During college, Yang worked in various research labs, including those specializing in plant molecular biology, plant physiology, and microbiology. She found her passion for the research of horticultural crops during her one-year study in Northwest A&F University, and started as a graduate student in the Department of Horticulture at Cornell University in the fall of 2009. Yang has been working with Dr. Xu on research projects investigating the regulation of fruit acidity in apple via genetic and genomic approaches. In 2012, she was awarded the Perrine Scholarship Award for her research in the field of pomology.

This dissertation is dedicated to my dear parents, Zhengming Bai and Wenjing Song, Mr. and Mrs. Paul Kisly, for their endless love and support

ACKNOWLEDGMENTS

I am tremendously grateful to my advisor, Dr. Kenong Xu, for his constant attention and guidance. Kenong is a wonderful mentor, whose expertise, understanding, and patience added considerably to my graduate experience. His guidance and encouragement have helped me grow beyond what I imagined was possible.

I am indebted to my special committee members: Dr. Lailiang Cheng and Dr. Jim Giovannoni for guiding the completion of my minor specializations. Lailiang and Jim are always encouraging, and their deep understanding in their areas and their dedication to excellence in both research and teaching are inspiring. The GC-MS metabolite profiling experiments in chapter 3 and 4 were conducted in the Cheng lab with the help of Lailiang himself and Ting Wu. My transcriptomic study was largely benefited from the helpful discussions with Jim himself, and with his group members: Dr. Silin, Zhong, Dr. Nigel Gapper and Yimin, Xu. I would like to gratefully appreciate Dr. Zhangjun Fei for his guidance in RNA-seq data analysis and statistical interpretation and Dr. Yi Zheng in the Fei lab for the pilot analysis of my data. In addition, I would like to sincerely thank Dr. Gennaro Fazio and his group, Dr. Ping Wang and his group, Dr. Judy Appleton and her group, Dr. Larry Smart and his group, Dr. Herb Aldwinkle, Dr. Susan Brown and Dr. Gan-Yuan Zhong for their generous help to my research.

I would like to thank the other members of the Xu lab for creating a wonderful working environment and being supportive all the time. In particular, I am grateful to Laura Dougherty, Dr. Dong Liang, Dr. Cuiying Li, Dr. Yuandi Zhu, Dr. Tuanhui, Bai and Dr. Aide Wang.

I also want to thank my extension supervisor Craig Cramer for his valuable guidance in writing and full support and understanding to my research.

I could have not got through this graduate school adventure without many wonderful friends and my family. Thanks to my parents, Zhengming Bai and Wenjing Song, and special thanks to Paul and Annie Kisly for their constant and unconditional support.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

A NATURAL MUTATION-LED TRUNCATION IN ONE OF THE TWO ALUMINUM-ACTIVATED MALATE TRANSPORTER-LIKE GENES AT THE *MA* LOCUS IS ASSOCIATED WITH LOW FRUIT ACIDITY IN APPLE[1]

1.1 Abstract

Acidity levels greatly affect the taste and flavor of fruit, and consequently its market value. In mature apple fruit, malic acid is the predominant organic acid. Several studies have confirmed that the major quantitative trait locus *Ma* largely controls the variation of fruit acidity levels. The *Ma* locus has recently been defined in a region of 150 kb that contains 44 predicted genes on chromosome 16 in the Golden Delicious genome. In this study, we identified two aluminum-activated malate transporter like genes, designated *Ma1* and *Ma2*, as strong candidates of *Ma* by narrowing down the *Ma* locus to 65-82 kb containing 12-19 predicted genes depending on the haplotypes. The *Ma* haplotypes were determined by sequencing two bacterial artificial chromosome clones from G.41 (an apple rootstock of genotype *Mama*) that cover the two distinct haplotypes at the *Ma* locus. Gene expression profiling in 18 apple germplasm accessions suggested that *Ma1* is the major determinant at the *Ma* locus controlling fruit acidity as *Ma1* is expressed at a much higher level than *Ma2* and the *Ma1* expression is significantly correlated with fruit titratable acidity (R^2=0.4543,

[1] Yang Bai, Laura Dougherty, Mingjun Li, Gennaro Fazio, Lailiang Cheng, Kenong Xu (2012). A natural mutation-led truncation in one of the two aluminum-activated malate transporter-like genes at the *Ma* locus is associated with low fruit acidity in apple. Molecular Genetics and Genomics 287(8):663-678

P=0.0021). In the coding sequences of low acidity alleles of *Ma1* and *Ma2*, sequence variations at the amino acid level between Golden Delicious and G.41 were not detected. But the alleles for high acidity vary considerably between the two genotypes. The low acidity allele of *Ma1*, *Ma1-1455A,* is mainly characterized by a mutation at base 1455 in the open reading frame. The mutation leads to a premature stop codon that truncates the carboxyl terminus of Ma1-1455A by 84 amino acids compared with Ma1-1455G. A survey of 29 apple germplasm accessions using marker $CAPS_{1455}$ that targets the SNP_{1455} in *Ma1* showed that the $CAPS_{1455A}$ allele was associated completely with high pH and highly with low titratable acidity, suggesting that the natural mutation-led truncation is most likely responsible for the abolished function of *Ma* for low pH or high acidity in apple.

1.2 Introduction

Organic acids, many of which are intermediates in metabolic processes, play significant roles in fruit growth, maturation, ripening and softening. The level of organic acids greatly affects the taste and flavor of fruit, and consequently its market value. The major determinants of fruit acidity include malic acid, citric acid and tartaric acid. In mature apple fruit, malic acid is the predominant organic acid although other organic acids such as citric acid, fumaric acid and quinic acid are detectable (Zhang et al. 2010). Apple fruit vary widely in pH and titratable acidity (TA) levels. However, the acceptable range for dessert apple fruit is often measured within a range of 3.1-3.8 in pH or 3.0-10.0 mg/ml in TA, beyond either end of which, fruit acidity is either too high or too low for fresh consumption (Brown and Harvey 1971; Nybom 1959; Visser and Verhaegh 1978).

Inheritance of high pH or low TA in apple fruit was attributed to a recessive gene in early studies (Brown and Harvey 1971; Nybom 1959; Visser and Verhaegh 1978). The acidity locus was mapped to linkage group (LG) 16 and designated as *Ma* (*malic acid*), where *Ma* is noted for the dominant low pH or high acidity allele and *ma* for high pH or low acidity allele (Maliepaard et al. 1998). In other species, major genes or quantitative trait locus (QTL) similar to *Ma* in controlling fruit acidity include *acitric* in citrus (Fang et al. 1997), *SS* in pomegranate (Jalikop 2007) and *pH* in sweet melon (Lerceteau-Köhler et al. 2012), where low acidity is also inherited recessively. The major gene *D* in peach, however, acts differently with low acidity being dominant over high acidity (Boudehri et al. 2009) although both peach and apple are members of the Rosaceae family. In tomato, complex and multiple QTLs are reported in conditioning fruit acidity levels (Fulton et al. 2002).

The primary role of the *Ma* locus in determining fruit pH and TA in apple was also demonstrated in QTL studies as a major QTL was consistently detected on LG 16 (Kenis et al. 2008; Liebhard et al. 2003; Xu et al. 2011). In addition to the *Ma* QTL, multiple minor QTLs of significant effect on acidity were identified in these studies. Although the minor QTLs are less consistent, the notion that the *Ma* locus and minor QTLs collectively determine fruit acidity levels is widely accepted. Consistent with this notion, a recent report finds that a mixed model of a major gene and polygenes fits best in explaining the apple acidity variation in a complex breeding population among four models (mixed, Mendelian, polygenic and environmental) tested (Iwanami et al. 2012).

Malate metabolism in fruit cells may involve several pathways according to recent reviews (Beruter 2004; Sweetman et al. 2009). Malate synthesis is considered to occur locally in fruit. The primary path is glycolysis of hexoses derived from sucrose and/or sorbitol, which are imported from leaves, in the cytosol of parenchyma

cells of fruit. Depending upon developmental stages, pathways of photosynthesis in the chloroplast, the tricarboxylic acid (TCA) cycle in mitochondrion, and glyoxylae cycle in glyoxysome in fruit cells also appear to be important for malate synthesis. For degradation of malate, gluconeogenesis and the TCA cycle are likely the main pathways. It is possible that the various enzymes involved in malate synthesis and degradation, such as phosphoenolpyruvate carboxylase (PEPC), NADP-dependent malic enzyme (NADP-ME), and NAD-dependent malate dehydrogenase (NAD-MDH) and many others, may play a role in regulating malate metabolism in fruit cells, thus acidity of fruit. In addition, the vacuolar transporters, such as the vacuolar pumps, e.g. V-ATPase (Schumacher and Krebs 2010), tonoplast dicarboxylate transporter, e.g. AttDT (Emmerlich et al. 2003), and members of the aluminum-activated malate transporter1 (ALMT1) family proteins (Barbier-Brygoo et al. 2011), e.g. AtALMT9 (Kovermann et al. 2007) and AtALMT6 (Meyer et al. 2011), may also play critical roles in determining fruit acidity as they can regulate the malate accumulation in and release from the vacuole in plant cells.

In apple, the pattern of malate accumulation and degradation is similar in developing fruits of several high/medium acid varieties studied, i.e. malic acid level significantly increases in young fruit (around 4 weeks after full bloom) and then progressively decreases through maturity although the total content per fruit increases along with fruit development (Beruter 2004; Hulme and Wooltorton 1957; Ulrich 1970; Zhang et al. 2010). Several recent studies have attempted to identify candidate genes and/or enzymes that may be associated with the acidity variations in apple fruit. Using a cDNA-AFLP-based approach, a gene designated *Mal-DDNA* (DQ417661) of unknown function previously appeared to be associated with low acid in a population segregating for fruit acidity (Yao et al. 2007). Direct profiling of expression patterns and enzyme activities of genes putatively involved in malate metabolism, including

MdPEPC (EU315246, for PEPC), *MdcyME* (DQ280492, for NADP-ME) and *MdVHA-A* (EF128033, for subunit A of vacuolar H^+-ATPase), found that there were differences between low and high acid genotypes (Yao et al. 2009). Involvement of genes encoding NADP-ME (GD254910, degradation of malate) and NAD-MDH (GD254856, synthesis of malate) in malate accumulation and degradation was also reported in a cDNA microarray analysis of 1,536 genes (Soglio et al. 2009). Moreover, a gene encoding NAD-MDH (DQ221207) has been functionally demonstrated to be involved in malate synthesis in apple (Yao et al. 2011). Overall, these data suggest that the genes and/or enzymes studied above may contribute to the variation of fruit acidity.

However, a detailed analysis of a low acid variety Usterapfel and its high acid mutant (Beruter 1998; 2004) indicated that key enzymes in malic acid metabolism, PEPC, NAD-MDH and NADP-ME, may not play a key role in determining the difference in fruit acidity because there was no difference in the catalytic activity of these enzymes between the two contrasting genotypes. Examining the localities of these genes in the apple genome (Velasco et al. 2010) appeared to support that these enzymes and genes involved in malate metabolism may not be *Ma* because none of those studied above, including *Mal-DDNA*, is on chromosome 16 where the *Ma* gene resides.

To uncover the genes underlying *Ma*, we had defined the *Ma* locus to a region of 150 kb encompassing 44 predicted genes on chromosome 16 in the Golden Delicious genome in a previous study (Xu et al. 2011). In this study, we report the identification of two aluminum-activated malate transporter (ALMT)-like genes, *Ma1* and *Ma2*, as strong candidates of *Ma*. We show that the *Ma* region is reduced to a genomic segment of 65 kb containing 19 predicted genes in Golden Delicious by developing three new markers and analyzing two more populations. In two bacterial artificial chromosome (BAC) clones that are distinguishable with haplotype *ma* and

Ma from apple rootstock G.41, the *Ma* region harbors 12 predicted genes, including *Ma1* and *Ma2*, although it spans over 71 kb in haplotype *ma* and 82 kb in haplotype *Ma*. We further show that the expression of *Ma1* is significantly correlated with fruit acidity levels, whereas *Ma2* is expressed constantly at low levels across high and low acidity fruit. Finally, we show that a single nucleotide mutation in the open reading frame of *Ma1* that leads to truncation of Ma1 by 84 amino acids is perfectly associated with high pH and highly with low TA in 29 apple germplasm accessions studied.

1.3 Materials and Methods

1.3.1 Plant materials and fruit pH and TA evaluation

Four half-sib F_1 populations of interspecific crosses were used to further narrow down the *Ma* locus, namely GMAL 4590, GMAL 4592, GMAL 4595 and GMAL 4596 (Table 1.1). The seed parent of the four populations is Royal Gala (*Mama*), a widely grown apple cultivar (*Malus* × *domestica* Borkh.). The pollen parents are elite clones of *M. sieversii* (i.e. of fruit size close to cultivated apple) collected from Kazakhstan (Forsline et al. 2003), including PI 613971 (*Mama*), PI 613978 (*mama*), PI 613988 (*Mama*) and PI 613979 (*Mama*), respectively. *M. sieversii* has been proven to be the major progenitor species of *M.* × *domestica* (Velasco et al. 2010). The four F_1 populations were derived from controlled crosses made in 2002 and planted on their own seedling roots in 2004 in the USDA-ARS Apple Germplasm Repository, Geneva, NY, USA. Populations GMAL 4590 of 216 individuals and GMAL 4595 of 222 genotypes were used in a previous study (Xu et al. 2011), but 36 and 23 individuals that did not bear fruit in 2010 from the two crosses, respectively, were not included

Table 1.1 List of populations segregating for fruit pH

Population	Seed Parent	Pollen Parent	All genotypes	Fruiting genotypes	pH≤3.8 (*Ma_*)	pH≥3.9 (*mama*)	Ratio (*Ma_:mama*)	P(χ^2)
GMAL 4590[a]	Royal Gala (*Mama*)	PI 613971 (*Mama*)	216	190	143	47	3:1	0.9300 (0.007)
GMAL 4592	Royal Gala (*Mama*)	PI 613978 *(mama)*	155	133	82	51	1:1	0.0070 (7.226)
GMAL 4595[b]	Royal Gala (*Mama*)	PI 613988 *(Mama)*	222	213	157	56	3:1	0.3400 (0.189)
GMAL 4596	Royal Gala (*Mama*)	PI 613979 *(Mama)*	215	198	156	42	3:1	0.2200 (1.515)
Total			808	734	538	196		

Estimated by pH paper. But pH meter reads for the informative recombinants (Figure1.1-1.2) and the 190 genotypes of GMAL 4595 (measured in 2010) were used

[a] The fruiting genotypes include one set of 36 bearing fruit in 2011 and the other set of 154 in 2010

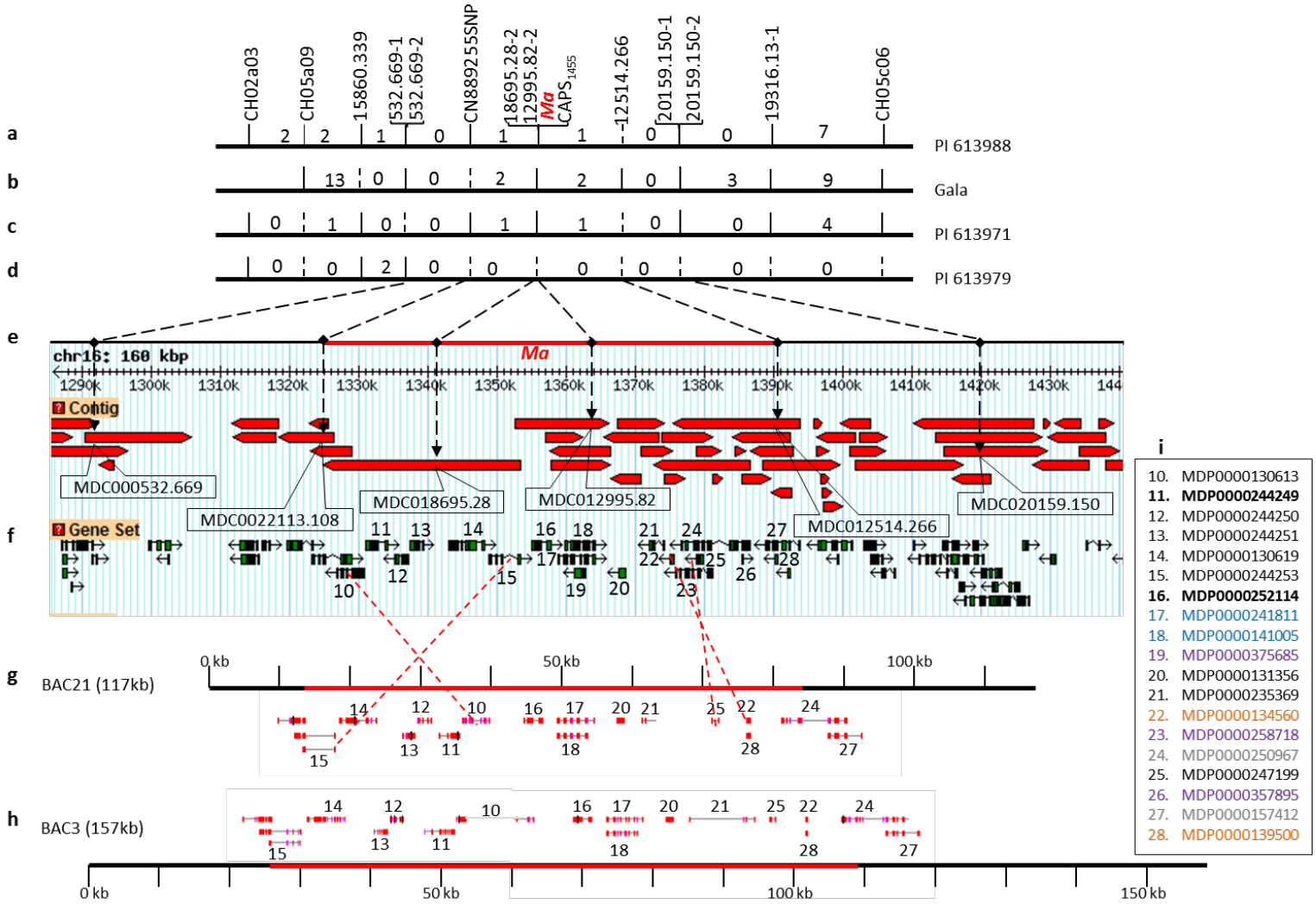
[b] The fruiting genotypes include one set of 23 bearing fruit in 2011 and the other of 190 in 2010

previously. These individuals bore fruit in 2011 and were added in this study (Table 1.1). Populations GMAL 4592 (155 genotypes) and GMAL 4596 (215 genotypes) were used for the first time. Overall, there are 724 fruiting individuals in a total of 808 genotypes in the four populations (Table 1.1).

Evaluation of fruit maturity, fruit acidity (pH paper estimates and instrumental measurements of pH and TA) was conducted similarly as described previously (Xu et al. 2011). Briefly, fruit maturity was determined via starch test that corresponds to Cornell Starch Index 4.0-6.0 (Blanpied and Silsby 1992). For pH estimates, pH paper (Hydrion Papers, pH 3.0-5.5, Micro Essential Laboratory Inc., Brooklyn, NY) was applied onto the fruit cuts at maturity in the orchard. For pH and TA instrumental measurements, fruit juice samples were prepared by pooling 5-10 fruits per genotype at maturity. The pooled juices were then measured with a pH meter (Accumet AB15, Fisher Scientific, Pittsburgh, PA), and subsequently, an autotitrator (Metrohm 848 Titrino Plus and Metrohm 869 Compact Sample Changer, Herisau, Switzerland). Evaluation for most genotypes was conducted either in 2010 or 2011. But for the informative recombinants between markers CH05c06 and CH02a03 or CH05a09 (Figures 1.1a-d, 1.2), pH meter based measurements were obtained in both years if fruit were available.

To examine the association between fruit acidity and the mutation at base 1455 in gene *Ma1*, pH and TA of mature fruit were evaluated for 29 representative apple cultivars and accessions (Table 1.2, including three progeny from GMAL 4595) grown in the USDA-ARS Apple Germplasm Repository, Geneva, New York.

Figure 1.1 Fine genetic and haploid specific physical maps of the *Ma* locus on chromosome 16. Fine genetic maps of *Ma* in PI 613988 (**a**), Royal Gala (**b**), PI 613971 (**c**) and PI 613979 (**d**) are shown. The number between the markers stands for the number of informative recombinants found in the interval. The solid vertical lines indicate the position of mapped markers, and the broken vertical lines are for positions of the presumed markers. **e** Physical map of the *Ma* region (a Genome Brower snapshot from the GDR website) in Golden Delicious (GD). The *Ma* region of 65 kb between markers CN889255SNP and 12514.266 is shown with a red solid bar. The labeled contigs indicate the source sequences, from which the markers were developed. **f** Predicted genes in the *Ma* region of GD. There are 19 predicted genes, which are conveniently labeled with #10-28, respectively. **g** A sequenced clone BAC21 of G.41 covering the *Ma* region. The numbers show the physical locations of the corresponding genes predicted in GD. **h** A sequenced clone BAC3 of G.41. **i** A list of the 19 genes predicted. Genes not present in G.41 (in purple): MDP0000375685 (#19), MDP0000258718 (#23) and MDP0000357895 (#26). Genes outside of the *Ma* region in G.41 (in grey): MDP0000250967 (#24) and MDP0000157412 (#27). Genes spliced alternatively (in blue): MDP0000241811 (#17) and MDP0000141005 (#18). Genes with duplicated IDs (in orange): MDP0000134560 (#22) and MDP0000139500 (#28). Candidate genes of *Ma* (in black): 12 genes, including MDP0000252114 (*Ma1*) and MDP0000244249 (*Ma2*)


CH02a03
CH05a09
15860.339
532.669-1
532.669-2
CN889255SNP
18695.28-2
12995.82-2
Ma
CAPS1455
12514.266
20159.150-1
20159.150-2
19316.13-1
CH05c06
a
b
c
d
e
f
g
h
i
PI 613988
Gala
PI 613971
PI 613979
chr16: 160 kbp
Ma
Contig
Gene Set
MDC000532.669
MDC0022113.108
MDC018695.28
MDC012995.82
MDC012514.266
MDC020159.150
BAC21 (117kb)
BAC3 (157kb)
0 kb
50 kb
100 kb
150 kb
10. MDP0000130613
11. MDP0000244249
12. MDP0000244250
13. MDP0000244251
14. MDP0000130619
15. MDP0000244253
16. MDP0000252114
17. MDP0000241811
18. MDP0000141005
19. MDP0000375685
20. MDP0000131356
21. MDP0000235369
22. MDP0000134560
23. MDP0000258718
24. MDP0000250967
25. MDP0000247199
26. MDP0000357895
27. MDP0000157412
28. MDP0000139500

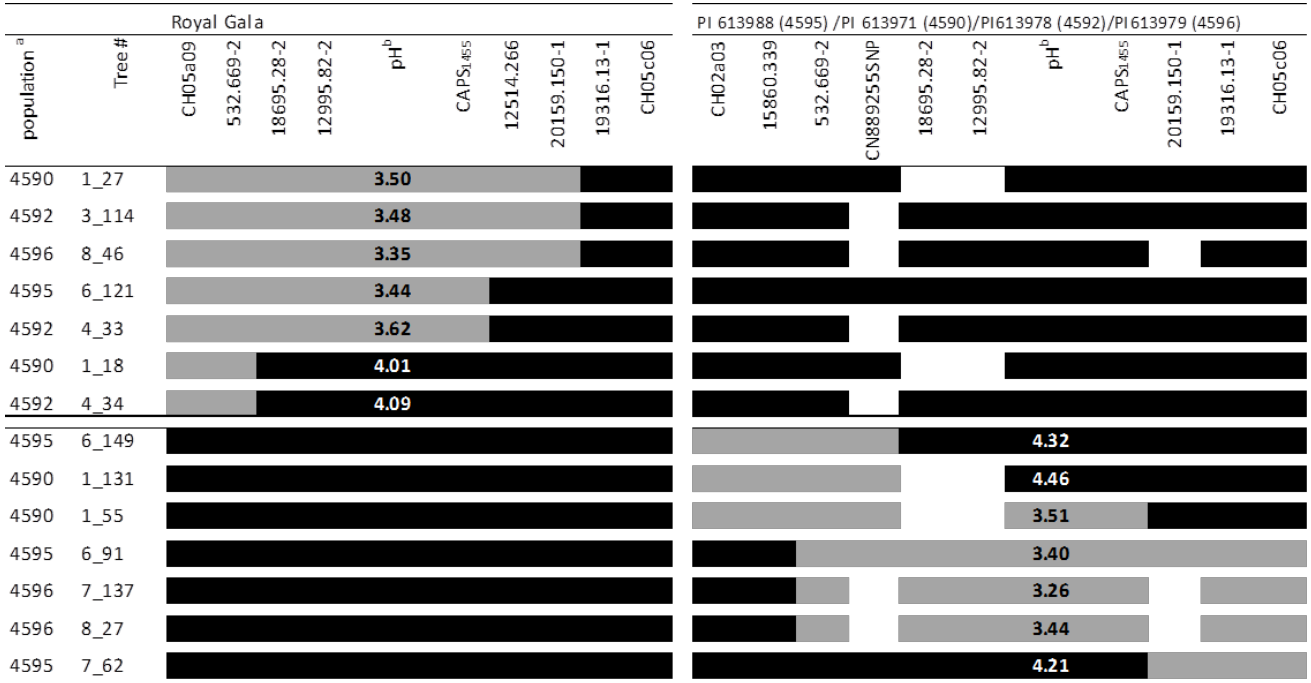


Figure 1.2 Key informative recombinants identified from populations GMAL 4590, 4592, 4595 and 4596 and their marker genotypes. [a] 4590=GMAL 4590; 4592=GMAL 4592; 4595=GMAL 4595; 4596=GMAL 4596. Marker genotype linked to the low pH (high acid) allele *Ma* in coupling phase is shown in gray, and those linked to the high pH (low acid) allele *ma* in black; [b] pH data for these recombinants were measured by a pH meter in 2010 (for GMAL 4595) and 2011 (for the rest)

Table 1.2 Apple germplasm accessions used for association survey between marker $CAPS_{1455}$ and fruit acidity. The 1st 18 apple accessions were used for qRT-PCR assay

Number	Name/accession #	Species
1	Britegold	*M. domestica*
2	Chisel Jersey	*M. domestica*
3	Cortland	*M. domestica*
4	Cox's Orange Pippin	*M. domestica*
5	E7-47	Hybrid
6	E7-54	Hybrid
7	Gala	*M. domestica*
8	Golden Delicious	*M. domestica*
9	Idared	*M. domestica*
10	Jonathan	*M. domestica*
11	KAZ 96 08-17	*M. Sieversii*
12	Marshall McIntosh	*M. domestica*
13	Monroe	*M. domestica*
14	Novosibirski Sweet	*M. domestica*
15	Poeltsamaa Winter Apple	*M. domestica*
16	PRI 1744-1	Hybrid
17	Sweet Delicious	*M. domestica*
18	Winter Majetin[a]	*M. domestica*
19	Delicious	*M. domestica*
20	Ein Shemer	*M. domestica*
21	PI 613988	*M. Sieversii*
22	Fuji	*M. domestica*
23	G.41	Rootstock, hybrid
24	Co-op 15	Hybrid
25	PI 613976	*M. Sieversii*
26	PI 613981	*M. Sieversii*
27	GMAL 4595-6-65	Hybrid
28	GMAL 4595-6-69	Hybrid
29	GMAL 4595-6-73	Hybrid

[a] Fruit were not matured about 30 day before maturity

1.3.2 Marker development and genetic mapping

New single sequence repeat (SSR) markers linked to *Ma* were developed using the same strategy as described in Xu et al (2011). Briefly, DNA sequences of contigs between the two existing markers 532.669-2 and 20159.150-1 (Xu et al. 2011) were downloaded from the Genome Database for Rosaceae (GDR, http://www.rosaceae.org/), and analyzed for the presence of potential SSRs markers using the web-based program BatchPrimer3 (http://probes.pw.usda.gov/batchprimer3/index.html) (You et al. 2008). Genomic DNA isolation, PCR and SSR analyses were conducted as described previously (Xu et al. 2011).

For single nucleotide polymorphism (SNP) marker development, we targeted expressed sequence tags (ESTs) that are present in the region between the two markers 532.669-2 and 20159.150-1. The presence and segregation of SNP were determined by direct sequencing of the PCR products amplified from the five parents and the informative recombinants between markers CH05c06 and CH02a03 or CH05a09.

$CAPS_{1455}$ is a cleaved amplified polymorphic sequences (CAPS) marker targeting base 1455 in the open reading frame of gene *Ma1*. The PCR program includes 2 min at 98 °C, 35 cycles of 10 s at 98 °C, 15 s at 55 °C and 90 s at 72 °C, and a final 5 min at 72 °C. PCR were conducted in a volume of 20 µl, which includes 1× PrimeSTAR® MAX DNA Polymerase (R045A, Takara/Clontech, Mountain View, CA), 0.5 mM of each primer and 30 ng of genomic DNA. Restriction digestion was performed at 37 °C for overnight in a volume of 20 µl that contains 10 µl PCR products, 2 U of BspHI (New England Biolabs, Ipswich, MA), 1× NEBuffer 4. Gel analysis of $CAPS_{1455}$ was conducted with agarose gels of 1.5% (w/v).

Mapping of markers in relation to the *Ma* locus was conducted with the informative recombinants between SSR markers Hi22f06 and CH02a03 or CH05a09,

which were identified from the four populations described above. The informative recombinants, as explained previously (Xu et al. 2011), refer to individual trees developed from zygotes that combined a parental-type gamete of an allele of *ma* (non-recombinant) with a recombined gamete near the *Ma* locus. Recombinants derived from zygotes that include a parental-type gamete of an allele of *Ma* (non-recombinant) are considered non-informative in this study. This is because the strong dominance effect of allele *Ma* from the parental-type gamete would make the effect of allele *Ma* or *ma* from a recombined gamete difficult, if not impossible, to detect by pH or TA values.

1.3.3 Identification of BAC clones and sequencing

The BAC library was constructed from G.41, an apple rootstock developed from an interspecific cross Malling 27 × Robusta 5 (Cummins et al. 2006). The mature fruit of G.41 are small (2-3 cm in diameter) and have astringent taste (not edible) and high acidity (pH=3.1, TA=13 mg/ml), suggesting that G.41 has a genotype of *MaMa* or *Mama*. The BAC library was constructed by Amplicon Express (Pullman, WA) using a restriction enzyme/vector combination of MboI/pECBAC1. It has a total of 41,472 clones with an average insert size of 120 kb, which provides approx. 6.6× coverage of the apple genome. The library was pooled at two levels with a total of five dimensions. The first level is the nine super pools, each of which comprises 12 plates containing a total of 4608 (12 × 384) clones. The second level is the nine sets of matrix pools, and each set was pooled from the 12 plates associated with one of the nine super pools. One set of matrix pool includes eight matrix plate pools (P:1.2.3; P:4.5.6; P:7.8.9; P:10.11.12; P:1.5.9; P:2.6.10; P:3.7.11; P:4.8.12) pooled from three of the 12 individual plate pools, eight matrix row pools (R:A.B.C.D; R:E.F.F.G; R:I.J.K.L;

R:M.N.O.P; R:A.E.I.M; R:B.F.J.N; R:C.G.K.O; R:D.H.L.P) pooled from four of the 16 individual row pools, ten matrix column pools (C:1.2.3.4.5.6.; C:7.8.9.10.11.12; C:13.14.15.16.17.18; C:19.20.21.22.23.24; C:1.7.13.19; C:2.8.14.20; C:3.9.15.21; C:4.10.16.22; C:5.11.17.23; C:6.12.18.24) pooled from six or four of the 24 individual column pools, and ten matrix diagonal pools (D:1.2.3.4.5.6; D.7.8.9.10.11.12; D.13.14.15.16.17.18; D.19.20.21.22.23.24; D:1.7.13.19; D:2.8.14.20; D:3.9.15.21; D:4.10.16.22; D:5.11.17.23; D:6.12.18.24) pooled from six or four of the 24 individual diagonal column pools. Different from a common individual column pool, which is pooled from the same column across a stack of 12 plates, an individual diagonal column pool comprises 12 varying columns on the diagonal line from the stack of 12 plates. For example, diagonal column pool D1 is pooled from column (C) 1 in plate (P) 1, C2 in P2… and C12 in P12, and pool D2 is from C2 in P1, C3 in P2… and C1 in P12.

Screening of BAC clones was conducted on the library super pools and their associated matrix pools using the *Ma*-linked PCR-based markers we developed. BAC clones originated from the *Ma* region were restricted with endonuclease BamHI and NotI (New England Biolabs, Ipswich, MA) and then analyzed by PFGE (pulse field gel electrophoresis) using CHEF-DR II System (Bio-Rad, Hercules, CA) for preliminary fingerprinting and size estimation. BAC sequencing was conducted using a 454 GS FLX system at Cornell Biotechnology Center and assembled with the Newbler Assembly (454 Life Sciences, Branford, CT).

1.3.4 Gene prediction and annotation at the Ma locus

Genes predicted in the *Ma* region of Golden Delicious (Velasco et al. 2010) were adopted and their CDS (coding sequences) and deduced protein sequences were

downloaded from GDR. Confirmation of gene annotation was carried out by searching the GenBank non-redundant protein database using the BLASTP program with a cutoff expected value of 10^{-9}. Putative functions of the predicted genes were annotated with the GenBank accession numbers of the highest similarities and associated functions if known.

1.3.5 Quantitative (q) RT-PCR assay of Ma candidate genes

Total RNA from mature fruit of 18 of the 29 apple accessions (Table 1.2) was isolated using Spectrum™ Plant Total RNA Kit (Sigma-Aldrich, St. Louis, MO) with three biological replicates. Reverse transcription reactions were carried out with 1.8 μg of total RNA using the Superscript III RT (Invitrogen, Carlsbad, CA). The resulting first strand cDNA was diluted by fivefold, and then used as templates for qRT-PCR analysis, in which a *Malus* (Gala) actin gene/EST (EB136338) served as a reference with primers Actin F (5'-GGCTGGATTTGCTGGTGATG-3') and Actin R (5'-TGCTCACTATGCCGTGCTCA-3').

Two rounds of qRT-PCR were performed. In the initial round, all 12 genes predicted at the *Ma* locus were screened with their gene specific primers (Table 1.3). Four low acid (Britegold, KAZ 96 08-17, Novosibirski Sweet and Sweet Delicious) and four high acid (Cox's Orange Pippin, Golden Delicious, Marshall McIntosh and Winter Majetin) apple accessions were used. cDNA of each genotype was bulked evenly from the three replicates and then used for qRT-PCR. In the second round, three selected genes (*Ma1*, *Ma2* and *MDP0000141005*) were analyzed in detail with all 18 apple accessions. The gene specific primers (Table 1.3) for *Ma1* are: Ma1F (5'-CGTCATGGTGTCTGGAACAT-3') and Ma1R (5'-CTCCATGGCAAAAACCTG

TC-3'), and those for Ma2 are Ma2F (5'-TCGGAAGACGGCCTAATGGA-3') and Ma2R (5'-TTGAAGCCGGGCAACAAACT-3'). These gene specific primers were designed to cover the known alleles of *Ma1* and *Ma2* (Figures 1.3, 1.4).

Table 1.3 qRT-PCR primers and the targeted genes at the *Ma* locus

Gene #	Gene Name	Forward and reverse primer sequence (5' to 3')	Expected PCR product size (bp)
10	MDP0000130613	AGAGGCGCAAGGACAAGCAC	184
		TTCACCACTGGCCAGCATTC	
11	**MDP0000244249 (*Ma2*)**	TCGGAAGACGGCCTAATGGA	200
		TTGAAGCCGGGCAACAAACT	
12	MDP0000244250	CATGTGCGGCTTCAATTTGG	203
		CGTTAATTTGGCGCCGAGAG	
13	MDP0000244251	CTTCGCGTGGAGCATGTGAT	207
		AACCGCGCGAATCTCTTGAA	
14	MDP0000130619	TGAAGGAGCTGTTGCCGTTG	195
		CCCTTCCCATGAGCAGCAGT	
15	MDP0000244253	TCTCCCGGAACCAGTCTTGC	206
		ACTTTGGCTGCACCGGGATA	
16	**MDP0000252114 (*Ma1*)**	CGTCATGGTGTCTGGAACAT	499
		CTCCATGGCAAAAACCTGTC	
17	MDP0000141005	TCTTGGCCGTTGAGGGCTGT	203
		CCGGAACCAGGTCCGTCCTC	
20	MDP0000131356	TACGTGCCCAAATGCCAAGA	194
		GCAGCTGGGAGAGGGTGTTC	
21	MDP0000235369	GGTCGAGACGTTGGCTACGC	223
		CCCCCATAAACTCGGCACAA	
22	MDP0000134560	GCCAAGGAAAATGCTTGGAGA	196
		GGTTCCCAACTTCCCGGTGT	
25	MDP0000247199	TGCTTGACGTGTTGGTTTCCA	194
		GTGCAGGGCTTCTCCTCCAA	
EB136338[a]	Actin (Ref.)	GGCTGGATTTGCTGGTGATG	179
		TGCTCACTATGCCGTGCTCA	

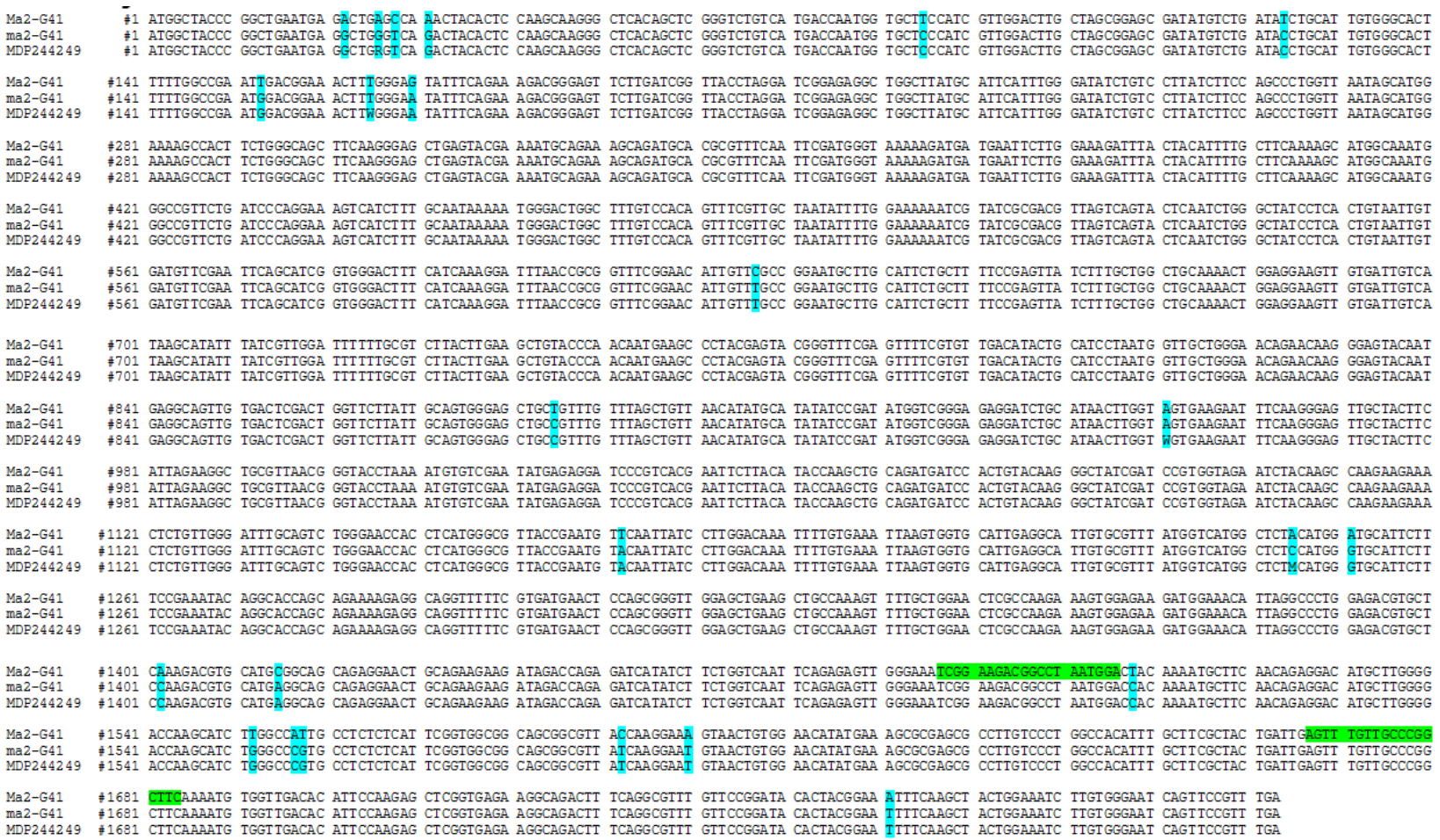
[a] a Gala actin gene/EST used as a reference in qRT-PCR

qRT-PCR was conducted using Roche (Indianapolis, IN) LightCycler 480 Real-Time PCR System. For each qRT-PCR, a final volume of 20 µl was used, which

Figure 1.3 Alignment of the *Ma1* allele coding sequences. *MDP252114* stands for the Golden Delicious gene *MDP0000252114*, which is a consensus sequence of alleles *Ma1-GD* and *ma1-GD*. SNPs are highlighted in blue. SNP_{1455A} leads to a stop codon TGA in allele *ma1-G41* as well as in *ma1-GD* as indicated by base R_{1455}, where R=G or A. Primer sites for qRT-PCR assay of *Ma1* are highlighted in green


Ma1-G41 #1 ATGGCGGCCA AAATCGGGTC CTTCCGCCAC AGCTTCGCCG AGAGAAGCAA GGAACGGCTG CTGTCGCGAA AAGGGTACTC CGACTTGGGC TTCAACAGCT CCGAGGCAGG CGACGAGCAC GTCAAATGCG GCTGCTTCCG
ma1-G41 #1 ATGGCGGCCA AAATCGGGTC CTTCCGCCAC AGCTTCGCCG AGAGAAGCAA GGAACGGCTG CTGTCGCGAA AAGGGTACTC CGACTTGGGC TTCAACAGCT CCGAGGCCGG CGACGAGCAC GTCAAATGCG GCTGCTTCCG
MDP252114 #1 ATGGCGGCCA AAATCGGGTC CTTCCGCCAC AGCTTCGCCG AGAGAAGCAA GGAACGGCTG CTGTCGCGAA AAGGGTACTC CGACTTGGGC TTCAACAGCT CCGAGGCCGG CGACGAGMAC GTCAAATGCG GCTGCTTCCG
Ma1-G41 #141 AAGAATCTCC GATGGGTTCA ACAACTTCTG CAACAATTTC CAGGGCACTT TCATCAAGTT ATATCAAATG GGTCATTCGG ATCCTCGAAA AGCCATCTTT GCTATCAAGA TGGGGTTGAC GTTGGCGCTC GTGTCGCTGC
ma1-G41 #141 AAGAATCTCC GATGGGTTCA ACAACTTCTG CAACAATTTC CAGGGCACTT TCATCAAGTT ATATCAAATG GGTCATTCGG ATCCTCGAAA AGCCATCTTT GCTATCAAGA TGGGGTTGAC GTTGGCGCTC GTGTCGCTGC
MDP252114 #141 AAGAATCTCC GATGGGTTCA AMAACTTCTG CAACAATTTC CAGGGCACTT TCATCAAGTT ATATCAAATG GGTCATTCGG ATCCTCGAAA AGCCATCTTT GCTATCAAGA TGGGGTTGAC GTTGGCGCTC GTGTCGCTGC
Ma1-G41 #281 TAATATTTTT CAAAGTGCCG CTCAAAGATG TCAGCCAGTA TTCTATCTGG GCAATTCTCA CTGTCGTCGT CGTCTTTGAA TTCAGCGTAG GTGCAACCTT GAACAAGGGT TTTAATCGTG CTTTGGGGAC GTTATCAGCG
ma1-G41 #281 TAATATTTTT CAAAGTGCCG CTCAAAGATG TCAGCCAGTA TTCTATCTGG GCAATTCTCA CTGTCGTCGT CGTCTTTGAA TTCAGCGTAG GTGCAACCTT GAACAAGGGT TTTAATCGTG CTTTGGGGAC GTTATCAGCG
MDP252114 #281 TAATATTTTT CAAAGTGCCG CTCAAAGATG TCAGCCAGTA TTCTATCTGG GCAATTCTCA CTGTCGTCGT CGTCTTTGAA TTCAGCGTAG GTGCAACCTT GAACAAGGGT TTTAATCGTG CTTTGGGGAC GTTATCAGCG
Ma1-G41 #421 GGCGGGCTTT CTCTTGGGAT TGCAGAGTTA TCCGTATTGG CCGGAGATCT GCAGGAAGTT ATAATTGTAA TTAGTGTGTT TATAGCAGGA TTTTGTGCTA GTTACGCCAA GCTTTATCCG TCAATGAAGC CATATGAATA
ma1-G41 #421 GGCGGGCTTT CTCTTGGGAT TGCAGAGTTA TCCGTATTGG CCGGAGATCT GCAGGAAGTT ATAATTGTAA TTAGTGTGTT TATAGCAGGA TTTTGTGCTA GTTACGCCAA GCTTTATCCG TCAATGAAGC CATATGAATA
MDP252114 #421 GGCGGGCTTT CTCTTGGGAT TGCAGAGTTA TCCGTATTGG CCGGAGATCT GCAGGAAGTT ATAATTGTAA TTAGTGTGTT TATAGCAGGA TTTTGTGCTA GTTACGCCAA GCTTTATCCG TCAATGAAGC CATATGAATA
Ma1-G41 #561 CGGATTTCGG GTATTCTTGT TGACATATTG TATCGTCATG GTGTCTGGAA CATCATTATT CTTTCAGACA GCAATTTATC GATTGCTGCT TATTGCTGTC GGTGCCGCCA CTAGTTTGAT TGTAAATATA TTTATATACC
ma1-G41 #561 CGGATTTCGG GTATTCTTGT TGACATATTG TATCGTCATG GTGTCTGGAA CATCATTATT CTTTCAGACA GCAATTTATC GATTGCTGCT TATTGCTGTC GGTGCCGCCA CTAGTTTGAT TGTAAATATA TTTATATACC
MDP252114 #561 CGGATTTCGG GTATTCTTGT TGACATATTG TATCGTCATG GTGTCTGGAA CATCATTATT CTTTCAGACA GCAATTTATC GATTGCTGCT TATTGCTGTC GGTGCCGCCA CTAGTTTGAT TGTAAATATA TTTATATACC
Ma1-G41 #701 CTATCTGGTC AGGGGAAGAT CTCCATAAGC TGGTGGTGAA AAATTTCAGG GGTGTTGCTG CTTCTTTGGA AGGTGTTGTT AATCAGTATC TGCAATGTGT TGAATACGAG AGAGTTCCTT CCAAAATTCT CACGTACCAA
ma1-G41 #701 CTATCTGGTC AGGGGAAGAT CTCCATAAGC TGGTGGTGAA AAATTTCAGG GGTGTTGCTG CTTCTTTGGA AGGTGTTGTT AATCAGTATC TGCAATGTGT TGAATACGAG AGAATTCCTT CCAAAATTCT CACGTACCAA
MDP252114 #701 CTATCTGGTC AGGGGAAGAT CTCCATAAGC TGGTGGTGAA AAATTTCAGG GGTGTTGCTG CTTCTTTGGA AGGTGTTGTT AATCAGTATC TGCAATGTGT TGAATACGAG AGAATTCCTT CCAAAATTCT CACKTACCAA
Ma1-G41 #841 GCTTCTGATG ACCCGTTGTA TAGTGGCTAT AGATCAGCTG TACAATCTTC AAGCGAAGAG GAAACTCTGT TAGGCTTTGC TGTATGGGAG CCACCCCATG GCCCATATAA GTCATTCAAT TATCCCTGGG TACACTATGT
ma1-G41 #841 GCTTCTGATG ACCCGTTGTA TAGTGGCTAT AGATCAGCTG TACAATCTTC AAGCGAAGAG GAAACTCTGT TAGGCTTTGC TGTATGGGAG CCACCCCATG GCCCATATAA GTCATTCAAT TATCCCTGGG TACACTATGT
MDP252114 #841 GCTTCTGATG ACCCGTTGTA TAGTGGCTAT AGATCAGCTG TACAATCTTC AAGCGAAGAG GAAACTCTGT TAGGCTTTGC TGTATGGGAG CCACCCCATG GCCCATATAA GTCATTCAAT TATCCCTGGG TACACTATGT
Ma1-G41 #981 CAAAGTTGCC GGTTCACTGA GGCATTGTGC ATTCATGGTC ATGGCGATGC ATGGATGCAT ACTTTCAGAA ATCCAGGCAC CAGCCGAAAA GAGACAGGTT TTTGCCATGG AGCTCCAGAG AGTTGGTGTG GAATGTGCTA
ma1-G41 #981 CAAAGTTGCC GGTTCACTGA GGCATTGTGC TTTCATGGTC ATGGCGATGC ACGGATGCAT ACTTTCAGAA ATCCAGGCAC CAGCCGAAAA GAGACAGGTT TTTGCCATGG AGCTCCAGAG AGTTGGTGTG GAATGTGCTA
MDP252114 #981 CAAAGTTGCC GGTTCACTGA GGCATTGTGC WTTCATGGTC ATGGCGATGC ACGGATGCAT ACTTTCAGAA ATCCAGGCAC CAGCCGAAAA GAGACAGGTT TTTGCCATGG AGCTCCAGAG AGTTGGTGTG GAATGTGCTA
Ma1-G41 #1121 AAATATTACG TGAGCTAGGA AGCAAAGTAG AAAAGATGGA AAAGTTAAGT CCCAAAGACG TACTTTTCGA TGTGCATGAA GCAGCAGAAG AGTTGCAGAT GAAAATAGAC AAATATTCTT ACCTTCTTGT CAATTCAGAG
ma1-G41 #1121 AAATATTACG TGAGCTAGGA AGCAAAGTAG AAAAGATGGA AAAGTTAAGT CCCAAAGACG TACTTTTCGA TGTGCATGAA GCAGCAGAAG AGTTGCAGAT GAAAATAGAC AAATATTCTT ACCTTCTTGT CAATTCAGAG
MDP252114 #1121 AAATATTACG TGAGCTAGGA AGCAAAGTAG AAAAGATGGA AAAGTTAAGT CCCAAAGACG TACTTTTCGA TGTGCATGAA GCAGCAGAAG AGTTGCAGAT GAAAATAGAC AAATATTCTT ACCTTCTTGT CAATTCAGAG
Ma1-G41 #1261 AGCTGGGGAC CTGCAGTACG TCCCAAGGAA TACGAAGATC ATGTGCATTT TGTCGATAAA GACAACGAAA ATAAACAGGT GGTGATCGAT TCTCTCAGTG AATATTGGGA TTCTCAGAAT CCAAGCACGA CTGTTGATCC
ma1-G41 #1261 AGCTGGGGAC CTGCAGTACG TCCCAGGGAA TACGAAGATC ATGTGCATTT TGTCGATAAA GACAACGAAA ATAAACAGGT GGTGATCGAT TCTCTCAGTG AATATTGGGA TTCTCAGAAT CCAAGCACGA CTGCTGATCC
MDP252114 #1261 AGCTGGGGAC CTGCAGTACG TCCCARGGAA TACGAAGATC ATGKGCATTT TGTCGATAAA GACAACGAAA ATAAACAGGT GGTGATCGAT TCTCTCAGTG AATATTGGGA TTCTCAGAAT CCAAGCACGA CTGCTGATCC
Stop codon SNP
Ma1-G41 #1401 TTCCAATCAA CAATGGATAT CTACAGAAAG TCTTTTAAAA CGACCAATAT CATGGCCACG CCTCTCATTC AACGCGCACG CTGTGCAGCA GGAACCTGAA GAATCAAAAG TTTACGAAAG CGCCAGCTCC TTATCTTTGG
ma1-G41 #1401 TTCCAATCAA CAATGGATAT CTACAGAAAG TCTTTTAAAA CGACCAATAT CATGACCACG CCTCTCATTC AACGCGCACG CTGTGCAGCA GGAACCTGAA GAATCAAAAG TTTACGAAAG CGCCAGCTCC TTATCTTTGG
MDP252114 #1401 TTCCAATCAA CAATGGATAT CTACAGAAAG TCTTTTAAAA CGACCAATAT CATGRCCACG CCTCTCATTC AACGCGCACG CTGTGCAGCA GGAACCTGAA GAATCAAAAG TTTACGAAAG CGCCAGCTCC TTATCTTTGG
Ma1-G41 #1541 CAACATTTGC ATCGCTTCTG ATTGAGTTTG TGGCTCGGCT TCAAAATCTG GTGGATGAAT TCGAAGAGCT GAGTGAGAAG GCAAATTTTA AGGATCCTGT TGAAGCCAAG GAGGAGGTTG TCGGGTTCTG GACCAAACTA
ma1-G41 #1541 CAACATTTGC ATCGCTTCTG ATTGAGTTTG TGGCTCGGCT TCAAAATCTG GTGGATGAAT TCGAAGAGCT GAGTGAGAAG GCAAATTTTA AGGATCCTGT TGAAGCCAAG GAGGAGGTTG TCGGGTTCTG GACCAAACTA
MDP252114 #1541 CAACATTTGC ATCGCTTCTG ATTGAGTTTG TGGCTCGGCT TCAAAATCTG GTGGATGAAT TCGAAGAGCT GAGTGAGAAG GCAAATTTTA AGGATCCTGT TGAACCCAAG GAGGAGGTTG TCGGGTTCTG GACCAAACTA
Ma1-G41 #1681 TTTAGGACTT TGCGGTTGAA GAACTAA
ma1-G41 #1681 TTTAGGAGTT TGCGGTTGAA GAACTAA
MDP252114 #1681 TTTAGGAGTT TGCGGTTGAA GAACTAA

Figure 1.4 Alignment of the *Ma2* allele coding sequences. MDP244249 stands for the Golden Delicious gene MDP0000244249, which is a consensus sequence of alleles Ma2-GD and ma2-GD. SNPs are highlighted in blue. Primer sites for qRT-PCR assay of Ma2 are highlighted in green


Ma2-G41 #1 ATGGCTACCC GGCTGAATGA GACTGAGCCA AACTACACTC CAAGCAAGGG CTCACAGCTC GGGTCTGTCA TGACCAATGG TGCTTCCATC GTTGGACTTG CTAGCGGAGC GATATGTCTG ATATCTGCAT TGTGGGCACT
ma2-G41 #1 ATGGCTACCC GGCTGAATGA GGCTGGGTCA GACTACACTC CAAGCAAGGG CTCACAGCTC GGGTCTGTCA TGACCAATGG TGCTCCCATC GTTGGACTTG CTAGCGGAGC GATATGTCTG ATACCTGCAT TGTGGGCACT
MDP244249 #1 ATGGCTACCC GGCTGAATGA GGCTGRGTCA GACTACACTC CAAGCAAGGG CTCACAGCTC GGGTCTGTCA TGACCAATGG TGCTCCCATC GTTGGACTTG CTAGCGGAGC GATATGTCTG ATACCTGCAT TGTGGGCACT
Ma2-G41 #141 TTTTGGCCGA ATTGACGGAA ACTTTGGGAG TATTTCAGAA AGACGGGAGT TCTTGATCGG TTACCTAGGA TCGGAGAGGC TGGCTTATGC ATTCATTTGG GATATCTGTC CTTATCTTCC AGCCCTGGTT AATAGCATGG
ma2-G41 #141 TTTTGGCCGA ATGGACGGAA ACTTTGGGAA TATTTCAGAA AGACGGGAGT TCTTGATCGG TTACCTAGGA TCGGAGAGGC TGGCTTATGC ATTCATTTGG GATATCTGTC CTTATCTTCC AGCCCTGGTT AATAGCATGG
MDP244249 #141 TTTTGGCCGA ATGGACGGAA ACTTWGGGAA TATTTCAGAA AGACGGGAGT TCTTGATCGG TTACCTAGGA TCGGAGAGGC TGGCTTATGC ATTCATTTGG GATATCTGTC CTTATCTTCC AGCCCTGGTT AATAGCATGG
Ma2-G41 #281 AAAAGCCACT TCTGGGCAGC TTCAAGGGAG CTGAGTACGA AAATGCAGAA AGCAGATGCA CGCGTTTCAA TTCGATGGGT AAAAAGATGA TGAATTCTTG GAAAGATTTA CTACATTTTG CTTCAAAAGC ATGGCAAATG
ma2-G41 #281 AAAAGCCACT TCTGGGCAGC TTCAAGGGAG CTGAGTACGA AAATGCAGAA AGCAGATGCA CGCGTTTCAA TTCGATGGGT AAAAAGATGA TGAATTCTTG GAAAGATTTA CTACATTTTG CTTCAAAAGC ATGGCAAATG
MDP244249 #281 AAAAGCCACT TCTGGGCAGC TTCAAGGGAG CTGAGTACGA AAATGCAGAA AGCAGATGCA CGCGTTTCAA TTCGATGGGT AAAAAGATGA TGAATTCTTG GAAAGATTTA CTACATTTTG CTTCAAAAGC ATGGCAAATG
Ma2-G41 #421 GGCCGTTCTG ATCCCAGGAA AGTCATCTTT GCAATAAAAA TGGGACTGGC TTTGTCCACA GTTTCGTTGC TAATATTTTG GAAAAAATCG TATCGCGACG TTAGTCAGTA CTCAATCTGG GCTATCCTCA CTGTAATTGT
ma2-G41 #421 GGCCGTTCTG ATCCCAGGAA AGTCATCTTT GCAATAAAAA TGGGACTGGC TTTGTCCACA GTTTCGTTGC TAATATTTTG GAAAAAATCG TATCGCGACG TTAGTCAGTA CTCAATCTGG GCTATCCTCA CTGTAATTGT
MDP244249 #421 GGCCGTTCTG ATCCCAGGAA AGTCATCTTT GCAATAAAAA TGGGACTGGC TTTGTCCACA GTTTCGTTGC TAATATTTTG GAAAAAATCG TATCGCGACG TTAGTCAGTA CTCAATCTGG GCTATCCTCA CTGTAATTGT
Ma2-G41 #561 GATGTTCGAA TTCAGCATCG GTGGGACTTT CATCAAAGGA TTTAACCGCG GTTTCGGAAC ATTGTTCGCC GGAATGCTTG CATTCTGCTT TTCCGAGTTA TCTTTGCTGG CTGCAAAACT GGAGGAAGTT GTGATTGTCA
ma2-G41 #561 GATGTTCGAA TTCAGCATCG GTGGGACTTT CATCAAAGGA TTTAACCGCG GTTTCGGAAC ATTGTTTGCC GGAATGCTTG CATTCTGCTT TTCCGAGTTA TCTTTGCTGG CTGCAAAACT GGAGGAAGTT GTGATTGTCA
MDP244249 #561 GATGTTCGAA TTCAGCATCG GTGGGACTTT CATCAAAGGA TTTAACCGCG GTTTCGGAAC ATTGTTTGCC GGAATGCTTG CATTCTGCTT TTCCGAGTTA TCTTTGCTGG CTGCAAAACT GGAGGAAGTT GTGATTGTCA
Ma2-G41 #701 TAAGCATATT TATCGTTGGA TTTTTTGCGT CTTACTTGAA GCTGTACCCA ACAATGAAGC CCTACGAGTA CGGGTTTCGA GTTTTCGTGT TGACATACTG CATCCTAATG GTTGCTGGGA ACAGAACAAG GGAGTACAAT
ma2-G41 #701 TAAGCATATT TATCGTTGGA TTTTTTGCGT CTTACTTGAA GCTGTACCCA ACAATGAAGC CCTACGAGTA CGGGTTTCGA GTTTTCGTGT TGACATACTG CATCCTAATG GTTGCTGGGA ACAGAACAAG GGAGTACAAT
MDP244249 #701 TAAGCATATT TATCGTTGGA TTTTTTGCGT CTTACTTGAA GCTGTACCCA ACAATGAAGC CCTACGAGTA CGGGTTTCGA GTTTTCGTGT TGACATACTG CATCCTAATG GTTGCTGGGA ACAGAACAAG GGAGTACAAT
Ma2-G41 #841 GAGGCAGTTG TGACTCGACT GGTTCTTATT GCAGTGGGAG CTGCTGTTTG TTTAGCTGTT AACATATGCA TATATCCGAT ATGGTCGGGA GAGGATCTGC ATAACTTGGT AGTGAAGAAT TTCAAGGGAG TTGCTACTTC
ma2-G41 #841 GAGGCAGTTG TGACTCGACT GGTTCTTATT GCAGTGGGAG CTGCCGTTTG TTTAGCTGTT AACATATGCA TATATCCGAT ATGGTCGGGA GAGGATCTGC ATAACTTGGT AGTGAAGAAT TTCAAGGGAG TTGCTACTTC
MDP244249 #841 GAGGCAGTTG TGACTCGACT GGTTCTTATT GCAGTGGGAG CTGCCGTTTG TTTAGCTGTT AACATATGCA TATATCCGAT ATGGTCGGGA GAGGATCTGC ATAACTTGGT WGTGAAGAAT TTCAAGGGAG TTGCTACTTC
Ma2-G41 #981 ATTAGAAGGC TGCGTTAACG GGTACCTAAA ATGTGTCGAA TATGAGAGGA TCCCGTCACG AATTCTTACA TACCAAGCTG CAGATGATCC ACTGTACAAG GGCTATCGAT CCGTGGTAGA ATCTACAAGC CAAGAAGAAA
ma2-G41 #981 ATTAGAAGGC TGCGTTAACG GGTACCTAAA ATGTGTCGAA TATGAGAGGA TCCCGTCACG AATTCTTACA TACCAAGCTG CAGATGATCC ACTGTACAAG GGCTATCGAT CCGTGGTAGA ATCTACAAGC CAAGAAGAAA
MDP244249 #981 ATTAGAAGGC TGCGTTAACG GGTACCTAAA ATGTGTCGAA TATGAGAGGA TCCCGTCACG AATTCTTACA TACCAAGCTG CAGATGATCC ACTGTACAAG GGCTATCGAT CCGTGGTAGA ATCTACAAGC CAAGAAGAAA
Ma2-G41 #1121 CTCTGTTGGG ATTTGCAGTC TGGGAACCAC CTCATGGGCG TTACCGAATG TTCAATTATC CTTGGACAAA TTTTGTGAAA TTAAGTGGTG CATTGAGGCA TTGTGCGTTT ATGGTCATGG CTCTACATGG ATGCATTCTT
ma2-G41 #1121 CTCTGTTGGG ATTTGCAGTC TGGGAACCAC CTCATGGGCG TTACCGAATG TACAATTATC CTTGGACAAA TTTTGTGAAA TTAAGTGGTG CATTGAGGCA TTGTGCGTTT ATGGTCATGG CTCTCCATGG GTGCATTCTT
MDP244249 #1121 CTCTGTTGGG ATTTGCAGTC TGGGAACCAC CTCATGGGCG TTACCGAATG TACAATTATC CTTGGACAAA TTTTGTGAAA TTAAGTGGTG CATTGAGGCA TTGTGCGTTT ATGGTCATGG CTCTMCATGG GTGCATTCTT
Ma2-G41 #1261 TCCGAAATAC AGGCACCAGC AGAAAAGAGG CAGGTTTTTC GTGATGAACT CCAGCGGGTT GGAGCTGAAG CTGCCAAAGT TTTGCTGGAA CTCGCCAAGA AAGTGGAGAA GATGGAAACA TTAGGCCCTG GAGACGTGCT
ma2-G41 #1261 TCCGAAATAC AGGCACCAGC AGAAAAGAGG CAGGTTTTTC GTGATGAACT CCAGCGGGTT GGAGCTGAAG CTGCCAAAGT TTTGCTGGAA CTCGCCAAGA AAGTGGAGAA GATGGAAACA TTAGGCCCTG GAGACGTGCT
MDP244249 #1261 TCCGAAATAC AGGCACCAGC AGAAAAGAGG CAGGTTTTTC GTGATGAACT CCAGCGGGTT GGAGCTGAAG CTGCCAAAGT TTTGCTGGAA CTCGCCAAGA AAGTGGAGAA GATGGAAACA TTAGGCCCTG GAGACGTGCT
Ma2-G41 #1401 CAAAGACGTG CATGCGGCAG CAGAGGAACT GCAGAAGAAG ATAGACCAGA GATCATATCT TCTGGTCAAT TCAGAGAGTT GGGAAATCGG AAGACGGCCT AATGGACTAC AAAATGCTTC AACAGAGGAC ATGCTTGGGG
ma2-G41 #1401 CCAAGACGTG CATGAGGCAG CAGAGGAACT GCAGAAGAAG ATAGACCAGA GATCATATCT TCTGGTCAAT TCAGAGAGTT GGGAAATCGG AAGACGGCCT AATGGACCAC AAAATGCTTC AACAGAGGAC ATGCTTGGGG
MDP244249 #1401 CCAAGACGTG CATGAGGCAG CAGAGGAACT GCAGAAGAAG ATAGACCAGA GATCATATCT TCTGGTCAAT TCAGAGAGTT GGGAAATCGG AAGACGGCCT AATGGACCAC AAAATGCTTC AACAGAGGAC ATGCTTGGGG
Ma2-G41 #1541 ACCAAGCATC TTGGCCATTG CCTCTCTCAT TCGGTGGCGG CAGCGGCGTT ACCAAGGAAA GTAACTGTGG AACATATGAA AGCGCGAGCG CCTTGTCCCT GGCCACATTT GCTTCGCTAC TGATTGAGTT TGTTGCCCGG
ma2-G41 #1541 ACCAAGCATC TGGGCCCGTG CCTCTCTCAT TCGGTGGCGG CAGCGGCGTT ATCAAGGAAT GTAACTGTGG AACATATGAA AGCGCGAGCG CCTTGTCCCT GGCCACATTT GCTTCGCTAC TGATTGAGTT TGTTGCCCGG
MDP244249 #1541 ACCAAGCATC TGGGCCCGTG CCTCTCTCAT TCGGTGGCGG CAGCGGCGTT ATCAAGGAAT GTAACTGTGG AACATATGAA AGCGCGAGCG CCTTGTCCCT GGCCACATTT GCTTCGCTAC TGATTGAGTT TGTTGCCCGG
Ma2-G41 #1681 CTTCAAAATG TGGTTGACAC ATTCCAAGAG CTCGGTGAGA AGGCAGACTT TCAGGCGTTT GTTCCGGATA CACTACGGAA ATTTCAAGCT ACTGGAAATC TTGTGGGAAT CAGTTCCGTT TGA
ma2-G41 #1681 CTTCAAAATG TGGTTGACAC ATTCCAAGAG CTCGGTGAGA AGGCAGACTT TCAGGCGTTT GTTCCGGATA CACTACGGAA TTTTCAAGCT ACTGGAAATC TTGTGGGAAT CAGTTCCGTT TGA
MDP244249 #1681 CTTCAAAATG TGGTTGACAC ATTCCAAGAG CTCGGTGAGA AGGCAGACTT TCAGGCGTTT GTTCCGGATA CACTACGGAA TTTTCAAGCT ACTGGAAATC TTGTGGGAAT CAGTTCCGTT TGA

contained 5 μl of the cDNA dilutions, 0.5 μM of the forward and reverse primers, and 1× SYBR Green Master Mix (Roche Cat. # 04707516001). The qRT-PCR program included an initial denaturation step of 10 min at 94 °C, 45 cycles of amplification using 10 s at 94 °C, 30 s at 58 °C, and 25 s at 72 °C, and a dissociation stage of 5 s at 95 °C, 60 s at 60 °C, and 15 s at 97 °C. Expression quantification and data analysis were performed by LightCycler 480 Software (Version 1.5) using the comparative cycle threshold method (Pfaffl 2001). Regression analysis between the gene expression and fruit acidity variation was performed using MS Excel 2007.

1.3.6 Phylogenetic analysis

Phylogenetic analysis of the deduced protein sequences of the *Ma* candidate genes, *Ma1* and *Ma2*, was conducted along with the members of the ALMT1 family in *Arabidopsis*, which sequences were downloaded from TAIR 10 (http://www.arabidopsis.org/index.jsp), using MEGA4 (Tamura et al. 2007).

1.4 Results

1.4.1 Delimiting the *Ma* locus to a 65 kb genomic segment on chromosome 16

Segregation of fruit pH in populations GMAL 4590 and GMAL 4595 had been studied previously, and the three parents Royal Gala, PI 613971 and PI 613988 had been determined of heterozygous genotype *Mama* (Xu et al. 2011). With additional fruiting individuals included, i.e. 36 in GMAL 4590 and 23 in GMAL 4595, the low pH (≤3.8) and high pH (≥3.9) segregation remained unchanged with the expected ratio

of 3:1 (Table 1.1). In population GMAL 4596, fruit pH segregated similarly with the ratio 3:1 (156:42, *P*=0.22), suggesting PI 613979, the pollen parent of GMAL 4596, is of a heterozygous genotype *Mama* as well. However, 82 low and 51 high pH genotypes were scored in population GMAL 4592, indicating a significant deviation from the 3:1 ratio (*P*=0.0004). Examining the markers linked to *Ma* (Figure 1.1) revealed that none of them segregated for the pollen parent PI 613978 while all segregated normally for Royal Gala. Moreover, the markers that segregate for Royal Gala alone predicted the segregation of pH (data not shown), suggesting that PI 613978 has a genotype of *mama*. Given the known genotype *Mama* of Royal Gala, pH is expected to segregate 1:1 in population GMAL 4592. But the observed ratio of 82:51 distorted significantly from 1:1 (*P*=0.007) (Table 1.1).

Three new markers, including two SSRs 12514.266 and 12995.82-2, and one SNP CN889255SNP, were developed (Figure 1.1a-d, Table 1.4) between the existing two markers 532.669-2 and 20159.150-1 that defined the *Ma* region previously (Xu et al. 2011). For map integration, the three new markers were assessed with a total of 52 informative recombinants between markers CH05c06 and CH02a03 or CH05a09, including 17 mapped in GMAL 4590, 7 in GMAL 4592, 19 in GMAL 4595, and 9 in GMAL 4596 (Figure 1.1a-d). Out of the 52 informative recombinants, 14 were the most informative in ordering the markers (Figure 1.2). SSR marker 12995.82-2 along with the existing marker 18695-28-2 cosegregated with *Ma*, and markers CN889255SNP and 12514.266 flanked *Ma* immediately to narrow the *Ma* locus down to a smaller genetic interval on chromosome 16 (Figures. 1.1a-d, 1.2). This genetic interval of *Ma* was supported with four most informative recombinants, including GMAL 4595-6-149 and GMAL 4590-1-131 between marker CN889255SNP and *Ma*, and GMAL 4592-4-33 and GMAL 4595-6-121 between *Ma* and marker 12514.266 (Figure 1.2). In physical terms, the *Ma* interval corresponds to a genomic segment of

Table 1.4 Primer sequences and other relevant information of markers developed in the *Ma* region

Marker Name	Marker Type	Primer-F/R (from 5' to 3')	IDs of source sequences	Targeted SSRs or bases	Expected size (bp)	Allele size (bp) or base for *Ma* [a]	Allele size (bp) or base for *ma* [b]	Detection	Genome mapped
CN889255SNP	SNP	GGAGGGTCTCCATCCAATTTA/ TCCCCACACATCTCATATTCC	CN889255/ MDC022113.108	base 330[c]	496(cDNA)/ 792(gDNA)	C	T	Sequencing	P2, P4
12995.82-2	SSR	AAAGCTTCTTCACACCAAGCA/ TGGTGATGATGGTGGTAGTCA	MDC012995.82	$(TC)_9$	173	164 (P1), 162 (P4, P5)	168 (P1-P5)	PAGE	P1,P4, P5
12514.266	SSR	AGGTATTGCCTAAATGTGTGTG/ TCACATCATAATGTTTTCCGAAT	MDC012514.266	$(TC)_{13}$	192	282 (P1)	290 (P1)	PAGE	P1
$CAPS_{1455}$	CAPS	GCCGCTTCTGGACTATCACTA/ TTCTTCAACCGCAAACTCCT	*Ma1*	TGG_{1455} /TGA_{1455}	2013	2013 (P1, P2, P4, P5)	1764+249 (P1-P5)	BspHI digestion + agarose gel electrophoresis	P1, P2, P4, P5

[a] Allele linked to *Ma* in coupling phase; [b] Allele linked to *ma* in coupling phas; [c] The 330th base from the first base of the fwd primer. The preceding sequence is AGGAATGGATTTTGGCTTCTAGGCTTGCAGCTTCTGATCAATTGGGTCT$(T/C)_{330}$. Genotype of Royal Gala is "T_{330}/T_{330}", and that of both PI 613971 and PI 613988 is "T_{330}/C_{330}". *P1* Royal Gala; *P2* PI 613971; *P3* PI 613978; *P4* PI 613988 and *P5* PI 613979; *PAGE* polyacrylamide gel electrophoresis

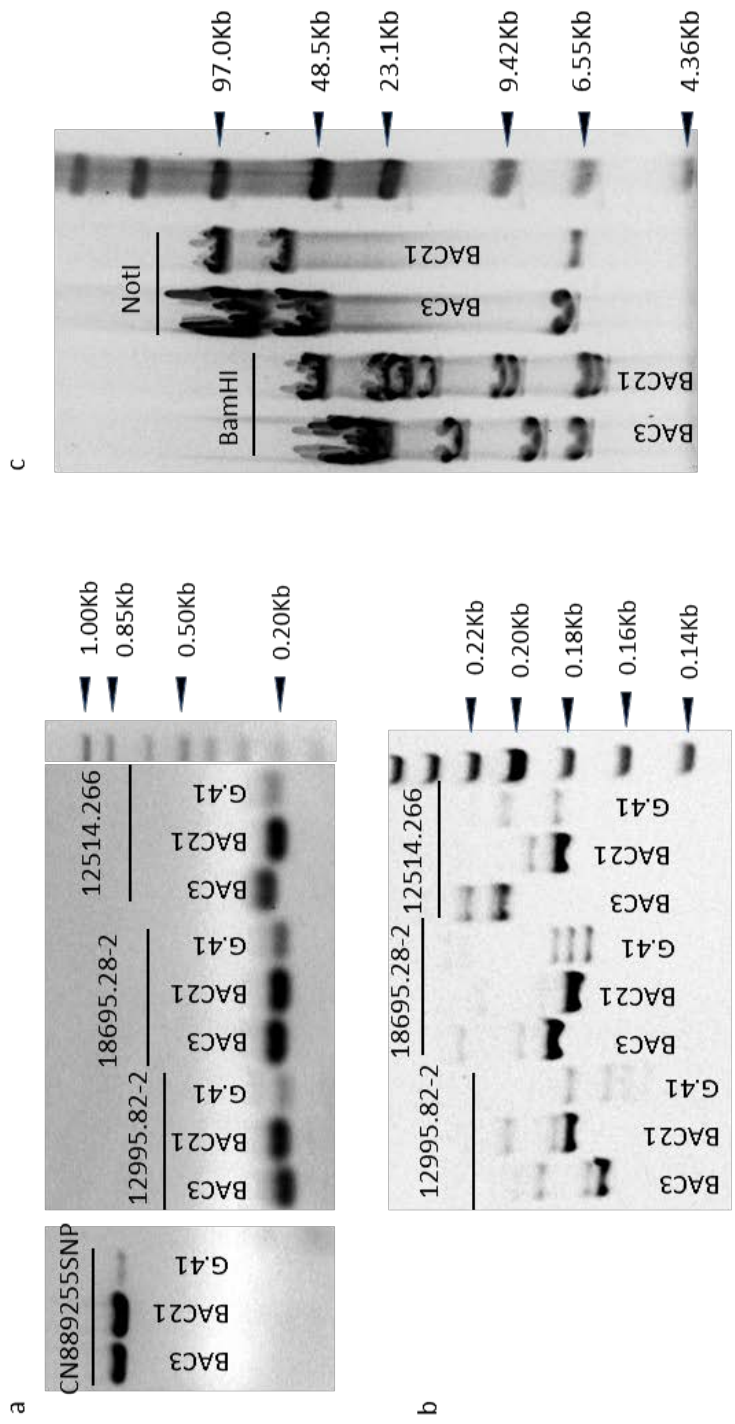
65 kb on chromosome 16 in Golden Delicious (Figure 1.1e), which was reduced from a 150 kb region defined previously (Xu et al. 2011).

1.4.2 Haplotypes of the Ma locus

The draft sequence of the Golden Delicious genome does not provide clear haplotype information although *M.* ×*domestica* is a highly heterozygous species. To understand the possible sequence variation and local genomic structure and organization that may discriminate allele *Ma* from *ma*, we identified two BAC clones BAC3 and BAC21 from the BAC library of apple rootstock G.41 using three markers 18695.28-2, 12995.82-2 and 12514.266 simultaneously. The two BAC clones were confirmed to contain not only the three makers used to screen the BAC library, but also the PCR amplicon source for marker CN889255SNP (Figure 1.5a), suggesting both BAC clones cover the *Ma* locus completely. Based on the band patterns associated with markers 18695.28-2, 12995.82-2 and 12514.266 (Figure 1.5b) and the restricted bands generated by BamHI and NotI digestions (Figure 1.5c), the two BACs are clearly of different haploid origin although the genotype of G.41 could be either *Mama* or *MaMa*. The estimated sizes for BAC3 and BAC21 were 150-160 kb and 110-120 kb, respectively (Figure 1.5c).

Sequencing of the two BAC clones revealed that the *Ma* region extends over a larger segment of 71 kb in BAC21 and 82 kb in BAC3 in G.41 (Figure 1.1g-h). Sequence alignment using BLAST demonstrated that BAC21 had higher overall sequence identity with the Golden Delicious contigs than BAC3 (data not shown), suggesting that BAC21 represents a haplotype likely closer to the two haplotypes in Golden Delicious than BAC3.

Figure 1.5 Identification and characterization of two BAC clones that completely cover the *Ma* locus and are of different haploid origin. **a** Agarose gel analysis of BAC clones BAC3 and BAC21 with markers CN889255SNP and 12514.266 (SSR) that delimit the *Ma* locus and markers (12995.82-2 and 18695.28-2) that co-segregate with *Ma* (Figure 1.1). G.41 genomic DNA was used as control; **b** Polyacrylamide gel analysis of BAC clones BAC3 and BAC21 with SSR markers 12995.82-2, 18695.28-2 and 12514.266 along with G.41 gDNA. Note that allele sizes for each marker vary but correspond to one of the bands amplified from G.41, **c** Pulse field gel electrophoresis of BAC3 and BAC21 digested with BamHI and NotI (1.0% agarose gel, run for 15 hrs at 6v/cm with switch time of 1-15 s)


a
CN889255SNP
BAC3
BAC21
G.41
12514.266
18695.28-2
12995.82-2
1.00Kb
0.85Kb
0.50Kb
0.20Kb
b
0.22Kb
0.20Kb
0.18Kb
0.16Kb
0.14Kb
c
NotI
BamHI
97.0Kb
48.5Kb
23.1Kb
9.42Kb
6.55Kb
4.36Kb

1.4.3 Identification of *Ma1* and *Ma2*

The 65 kb genomic region of *Ma* contains 19 predicted genes in Golden Delicious (#10-28 in the 44 genes listed in Xu et al. (2011), Figure 1.1f, i; Table 1.5). Aligning the 19 predicted genes with the two BACs indicated that three genes (*MDP0000375685* (#19), *MDP0000258718* (#23) and *MDP0000357895* (#26)) were not found in the two BAC sequences, and two (*MDP0000250967* (#24) and *MDP0000157412* (#27) reside outside of the *Ma* region defined by the two markers CN889255SNP and 12514.266 in G.41. Moreover, *MDP0000134560* (#22) and *MDP0000139500* (#28) are duplicated gene IDs for a single gene, and *MDP0000241811* (#17) and *MDP0000141005* (#18) are alternatively spliced variants from another single gene. Therefore, the two BAC clones harbor 12 predicted genes at the *Ma* locus (Figure 1.1g-i; Tables 1.3, 1.5), which include *MDP0000252114* (#16), designated *Ma1*, and *MDP0000244249* (#11), designated *Ma2*. Proteins Ma1 and Ma2 are putative members of the ALMT1 family and respectively share 57% (338/595) and 55% (302/553) of identity in amino acid sequence with AtALMT9, an *Arabidopsis* protein known to be a vacuolar malate channel involved in maintaining the cytosolic malate homeostasis (Kovermann et al. 2007). A search of *Malus* EST databases in GenBank found that there are 20 EST accessions of the origin of *Ma1* (Table 1.6) and one EST (CN929391) matching with *Ma2*, suggesting both *Ma1* and *Ma2* are expressed genes, and therefore strong candidate genes of *Ma*.

There are two inversions in gene orders between Golden Delicious and G.41 (Figure 1.1f,g): one between genes *MDP0000130613* (#10) and *MDP0000244253* (#15), and the other between genes *MDP0000134560* (#22) and *MDP0000247199*. Genes *Ma1* and *Ma2* are physically separated by one gene *MDP0000130613* (#10) in both haplotypes of G.41 (Figure 1.1g,h), but by four genes *MDP0000244250* (#12),

Table 1.5 List of genes predicted in the *Ma* region

Gene #	*Malus* Gene Id	Description in apple genome	Hit accession	Description of Hit accessions in GenBank	Max Score	e-value
10	MDP0000130613	type 1 domain, K Homology, K Homology domain, RNA binding molecular function Interacting selectively with an RNA molecule or a portion	XP_003519473	ribosomal RNA assembly protein mis3-like [Glycine max]	559	0
11	**MDP0000244249 (*Ma2*)**	Protein of unknown function UPF0005 family	XP_002278994	aluminum-activated malate transporter 9 [Vitis vinifera]	704	0
12	MDP0000244250	Tat binding protein 1-interacting family, or F-duction.	XP_002279040	homologous-pairing protein 2 homolog [Vitis vinifera]	327	1.05E-107
13	MDP0000244251	Protein of unknown function DUF292	XP_002529127	protein with unknown function [Ricinus communis]	387	1.53E-127
14	MDP0000130619	protein serine/threonine kinase activity molecular function Catalysis of the reaction: ATP + a protein serine/threonine	ABA99626	Protein kinase domain containing protein, expressed [Oryza sativa Japonica Group]	634.795	0
15	MDP0000244253	Heavy metal transport/detoxification protein domain, within or between cells.	XP_002276388	uncharacterized protein LOC100245724 [Vitis vinifera]	173	1.59E-49
16	**MDP0000252114 (*Ma1*)**	Protein of unknown function UPF0005 family	XP_002278978	aluminum-activated malate transporter 4 [Vitis vinifera]	790	0
17	MDP0000241811[d]	Armadillo-type fold domain, Armadillo-like helical domain, HEAT repeat, cytosol cellular component That part of the cytoplasm that does not contain membranous or particulate subcellular components. and consists of catalytic and regulatory subunits.	AEQ75494	serine/threonine protein phosphatase 2a regulatory subunit A [Rosa multiflora]	1153	0
18	MDP0000141005[d]	Armadillo-type fold domain, Armadillo-like helical domain, HEAT repeat, cytosol cellular component That part of the cytoplasm that does not contain membranous or particulate subcellular components. and consists of catalytic and regulatory subunits.	AEQ75494	serine/threonine protein phosphatase 2a regulatory subunit A [Rosa multiflora]	1147	0
19	MDP0000375685[a]	Myb-type HTH DNA-binding domain domain, Myb transcription factor family	ABH02906	MYB transcription factor MYB161 [Glycine max]	270	9.34E-86

Table 1.5 (Continued)

20	MDP0000131356	Alpha/beta hydrolase fold-3 domain, and protein synthesis and degradation.	XP_00227 8939	probable carboxylesterase 6-like [Vitis vinifera]	433	2.9E-147
21	MDP0000235369	UAG, usually in response to a termination codon (UAA, translational termination biological process	NP_00115 1538	eukaryotic peptide chain release factor subunit 1-1 [Zea mays]	78	1.638 E-13
22	MDP0000134560[c]	histone-lysine N-methyltransferase activity molecular function Catalysis of the reaction: S-adenosyl-L-methionine + histone L-lysinezinc ion binding molecular function Interacting selectively with zinc (Zn) ions.	XP_00354 1933	probable histone-lysine N-methyltransferase ATXR3-like [Glycine max]	200	1.62E -56
23	MDP0000258718[a]	branched-chain-amino-acid transaminase activity molecular function Catalysis of the reaction: L-leucine + 2-oxoglutarate	XP_00253 0599	branched-chain amino acid aminotransferase, putative [Ricinus communis]	622	0
24	MDP0000250967[b]	bZIP-1 domain, bZIP transcription factor, Protein of unknown function DUF580 family, Basic-leucine zipper (bZIP) transcription factor domain, or with a specific sequence motif or type of DNA e.g. promotor binding or rDNA binding.	XP_00228 2011	CTL-like protein DDB_G0274487 [Vitis vinifera]	710	0
25	MDP0000247199	Protein of unknown function Cys-rich family	XP_00363 1187	protein PLANT CADMIUM RESISTANCE 2-like [Vitis vinifera]	213	1.118 E-66
26	MDP0000357895[a]	No Description		No Hit		
27	MDP0000157412[b]	Protein of unknown function DUF580 family, many organelles may be a single or double lipid bilayer also includes associated proteins.	XP_00228 2011	CTL-like protein DDB_G0274487 [Vitis vinifera]	765	0
28	MDP0000139500[c]	histone-lysine N-methyltransferase activity molecular function Catalysis of the reaction: S-adenosyl-L-methionine + histone L-lysinezinc ion binding molecular function Interacting selectively with zinc (Zn) ions.	XP_00354 1933	probable histone-lysine N-methyltransferase ATXR3-like [Glycine max]	200	1.62E -56

[a] Not found in BAC3 and BAC21; [b] Beyond the *Ma* region in BAC3 and BAC21; [c] Identical genes; [d,...]Alternatively spliced genes

Table 1.6 *Ma1* derived EST accessions in GenBank

EST accession #	Source	Tissue type	TGG_{1455}/TGA_{1455}	Notes
CO723101	GoldRush	bud	TGG	
CX024250	GoldRush	leaf	TGG	
CN494439	GoldRush	flower	TGG	
GO547092	Granny	Fruit	TGG	
GO509271	*Malus* x domestica	flower	TGG	baloon stage
GO562003	Geneva 3041 x *Malus* sieversii	root	TGG	phytophtora challenged young root
HM641023	Green Sleeves	Fruit	TGG	
CN865511	Royal Gala	Fruit	NA	150 DAFB fruit cortex
CN876849	Royal Gala	leaf	NA	partially senescing leaf
CN907263	Royal Gala	leaf	NA	temperature stressed leaves
CO068148	GoldRush	Fruit	NA	
CN494439	GoldRush	flower	NA	
GO525361	Granny	Fruit	NA	
DT042003	Geneva 3041 x *Malus* sieversii	root	NA	phytophtora challenged young root
GO539195	Royal Gala	xylem	NA	plants subjected to 5 degrees C. for 24hrs
CN860595	Royal Gala	fruit	NA	150 DAF bloom, skin peel
CN860600	Royal Gala	fruit	NA	150 DAF bloom, skin peel
CN893992	Royal Gala	fruit	NA	126 DAFB, cortex
CN914864	Braeburn	fruit cells	NA	cultured fruit cells, boron exposed, phytoremediation.
CN917883	M9	root	NA	root tips (distal 1.5 cm)

MDP0000244251 (#13), *MDP0000130619* (#14) and *MDP0000244253* (#15) in Golden Delicious (Figure 1.1f).

1.44 qRT-PCR analysis of genes predicted at the Ma locus

To investigate the expression patterns in mature fruit, genes *Ma1* and *Ma2* as well as the other ten genes in the *Ma* region were screened alongside four low acid (Britegold, KAZ 96 08-17, Novosibirski Sweet and Sweet Delicious) and four high acid (Cox's Orange Pippin, Golden Delicious, Marshall McIntosh and Winter Majetin) apple germplasm accessions (Table 1.2) using qRT-PCR. Gene *Ma1* was expressed at much higher levels in high acid fruit than in low acid fruit while *Ma2* was expressed consistently at low levels across both low and high acid fruit (data not shown). The correlation between gene expression and TA among the eight apple accessions was highly significant for *Ma1* (R^2=0.9430, P=0.0001) but non-significant for *Ma2* (R^2=0.0559, P=0.5729). Among the other ten genes, *MDP0000141005*, which encodes a putative serine/threonine protein phosphatase 2A (PP2A) regulatory subunit A, was expressed at high levels and showed a significant correlation with fruit acidity (R^2=0.7428, P=0.0059). The remaining nine genes were expressed at low levels and did not show correlations with fruit acidity (data not shown), and therefore we did not analyze them further.

A more comprehensive qRT-PCR analysis of *Ma1*, *Ma2 and MDP0000141005* indicated that the relative expression levels of *Ma1* remained high and were significantly correlated with TA (R^2=0.4543, P=0.0021) and pH (R^2=0.4630, P=0.0019) in fruit of 18 apple germplasm accessions (Figure 1.6a, b). In contrast, the expression of *Ma2* was low and showed no correlation with TA (R^2=0.0086, P=0.7148) and pH (R^2=0.0356, P=0.4531) (Figure 1.6c, d). These data suggest that

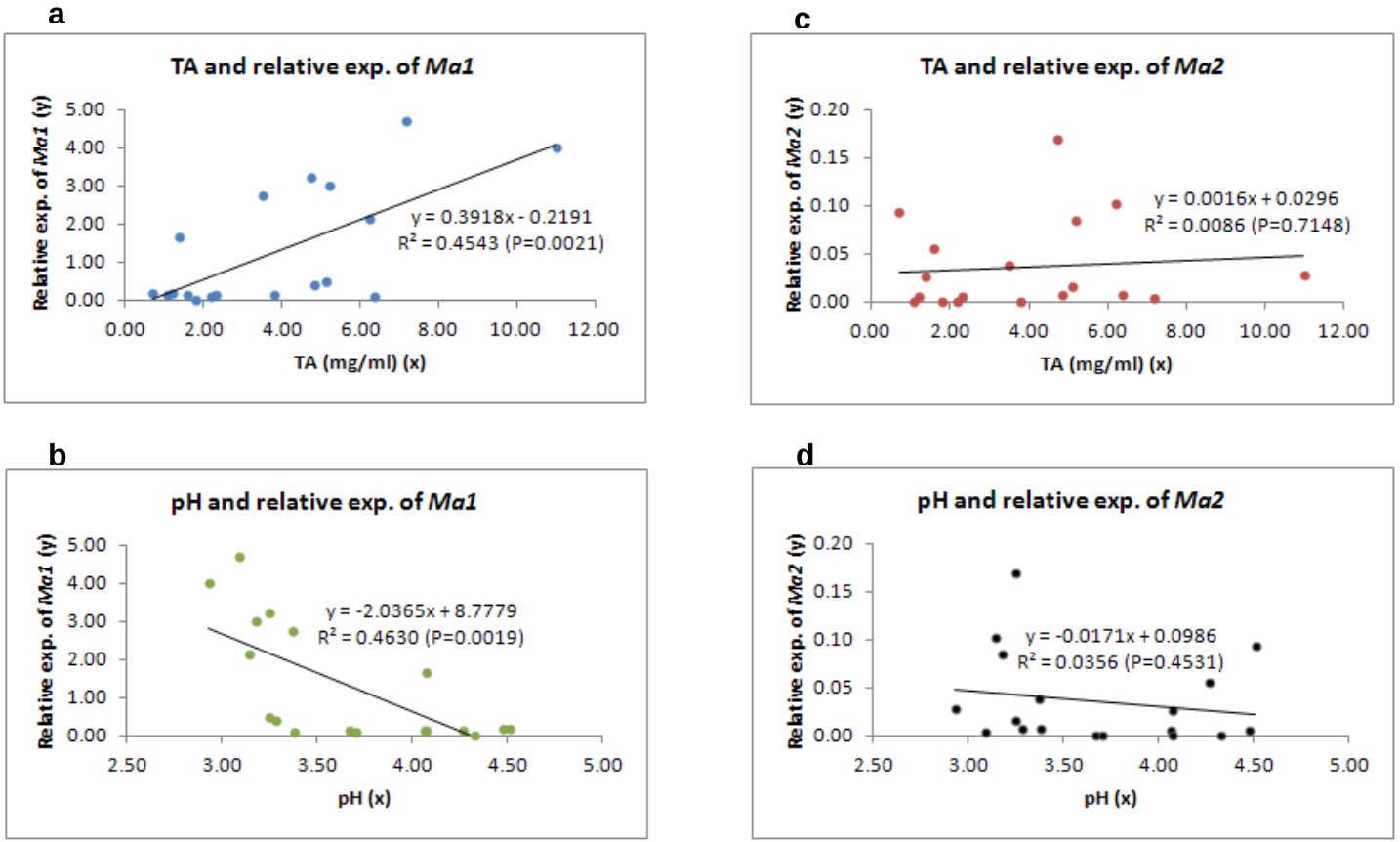


Figure 1.6 Regression between fruit acidity (TA and pH) and relative gene expression (*Ma1* and *Ma2*) in 18 apple germplasm accessions. **a-d** Self explainable

Ma1 may be the major factor in determining fruit acidity and the role of *Ma2* would be limited if any. The correlation of *MDP0000141005* expression with acidity was reduced to a non-significant level (R^2=0.1497, P=0.1126 for TA; R^2= 0.0916, P=0.2222 for pH), allowing *MDP0000141005* to be excluded from subsequent analyses.

1.45 Allelic variations of *Ma1* and *Ma2*

The *Ma1* allele in BAC3, designated *Ma1-G41*, differed by eight bases from that in BAC21, designated *ma1-G41* (Figure 1.3, Table 1.7). Examining the coding sequence of *MDP0000252114* showed that nucleotides at seven positions are ambiguous, i.e. M=A/C (bases 118 and 162); K=G/T (bases 834 and 1304); W=A/T (base 1011) and R=A/G (bases 1286 and 1455), presumably caused by the two different haplotypes in Golden Delicious (Figure 1.3, Table 1.7). To distinguish the two alleles of *Ma1* in Golden Delicious, we compared the sequence of MDP0000252114 with both *Ma1-G41* and *ma1-G41*. Excluding the seven ambiguous positions, *MDP0000252114* differed by one base from *ma1-G41*, but by six bases from *Ma1-G41*, suggesting *MDP0000252114* is much closer to *ma1-G41* than to *Ma1-G41*. When the seven ambiguous positions were considered, one of the two possible bases at each of the seven positions matches with the base at their corresponding positions in *ma1-G41*. This set of seven bases was therefore inferred to be co-present in one allele of Golden Delicious, designated *ma1-GD*. The other set of seven bases were concluded to be co-present in the other allele of Golden Delicious, designated *Ma1-GD* (Figure 1.3, Table 1.7).

In the deduced amino acid sequences, Ma1-G41 and Ma1-GD diverge by seven residues while there is no difference between ma1-G41 and ma1-GD

Table 1.7 DNA and amino acid sequence variations in the *Ma1* alleles of G.41and Golden Delicious (GD)

Base position[a]	DNA					AA				
	Ma1-G41[b]	*ma1-G41*[b]	*Ma1-GD*[c]	*ma1-GD*[c]	*MDP0000252114*[c]	*Ma1-G41*	*ma1-G41*	*Ma1-GD*	*ma1-GD*	*MDP0000252114*
108	A	C	C	C	C	A	A	A	A	A
118	C	C	A	C	M=A,C	H	H	N	H	H,N
162	C	C	A	C	M=A,C	N	N	K	N	N,K
814	G	A	A	A	A	V	I	I	I	I
834	G	G	T	G	K=G,T	T	T	T	T	T
1011	A	T	A	T	W=A,T	A	A	A	A	A
1032	T	C	C	C	C	H	H	H	H	H
1286	A	G	A	G	R=A,G	K	R	K	R	K,R
1304	T	T	G	T	K=G,T	V	V	G	V	V,G
1394	T	C	C	C	C	V	A	A	A	A
1455	G	A	G	A	R=A,G	W	STOP	W	STOP	W/STOP
1645	G	G	C	C	C	A	A	P	P	P
1688	C	G	G	G	G	T	S	S	S	S

[a] Counted from the 1st base in the coding sequences

[b] There are eight base variations in the coding sequence between *Ma1-G41* and *ma1-G41*. Out of the eight base variations, three are silent mutations and five are pronounced, including the one at the 1455th base that led to a stop codon in *ma1-G41* for premature termination. As a result, another pronounced mutation at base 1688 was beyond the coding sequence of *ma1-G41*

[c] There are seven ambiguous nucleotides in the coding sequence of MDP0000252114. The 1455th base is R, an ambiguous base for A or G, suggesting a similar premature stop codon in *ma1-GD*

(Figure 1.7, Table 1.7). However, both ma1-G41 and ma1-GD are truncated by 84 amino acids at the carboxyl terminus compared with either Ma1-G41 or Ma1-GD (Figure 1.7). This truncation is due to a nucleotide mutation from G to A at the 1455th base (SNP_{1455}) in the open reading frame, leading to a pronounced change from a tryptophan (W) codon TGG_{1455} to a stop codon TGA_{1455} (Figures 1.7,1.3; Table 1.7).

The allelic variations of gene *Ma2* were investigated similarly (Figures 1.4, 1.8, Table 1.8). Briefly, the *Ma2* allele in BAC3 and that in BAC21 were designated *Ma2-G41* and *ma2-G41*, respectively, whereas *Ma2-GD* and *ma2-GD* were assigned as two alleles for Golden Delicious based on the *MDP0000244249* sequence of four ambiguous positions, i.e. R=A/G (base 26), W=A/T (bases 165 and 951) and M=A/C (base 1245). There are 24 different bases (17 aa) between *Ma2-G41* and *ma2-G41*, 4 bases (2 aa) between *Ma2-G41* and *Ma2-GD*, and 22 bases (17 aa) between *Ma2-G41* and *ma2-GD*. The coding sequences in alleles *ma2-G41* and *ma2-GD* are identical (Figures 1.4, 1.8; Table 1.8).

1.46 Allelic association of the *Ma1* and *Ma2* alleles with *Ma* and *ma*

To uncover which *Ma1* and *Ma2* allele is associated with *Ma* or *ma*, a CAPS marker, named $CAPS_{1455}$, was developed to target SNP_{1455} using endonuclease BspHI, which cleaves site $TCATGA_{1455}$ in the truncated alleles *ma1-G41* or *ma1-GD*, but not $TCATGG_{1455}$ in the intact alleles *Ma1-G41* or *Ma1-GD* (Table 1.4). Agarose gel assay of maker $CAPS_{1455}$ in population GMAL 4595 and the informative recombinants indicated that homozygous genotype $CAPS_{1455G}CAPS_{1455G}$ cosegregated with *MaMa*, $CAPS_{1455G}CAPS_{1455A}$ with *Mama*, and $CAPS_{1455A}CAPS_{1455A}$ with *mama* (Figures 1.1, 1.2, 1.9a), suggesting the intact allele of *Ma1* (*Ma1-1455G*), such as *Ma1-G41* or *Ma1-GD*, is associated with the high acid allele *Ma* while the truncated allele of *Ma1*

Figure 1.7 Alignment of the Ma1 deduced protein sequences. MDP252114 stands for the Golden Delicious protein MDP0000252114, which combines proteins Ma1-GD and ma1-GD. Each sign "-" in MDP252114 is for two possible amino acid residues, and annotated accordingly as shown. Amino acid residues that vary are highlighted in blue. The stop codon TGA_{1455} caused by SNP_{1455A} leads to a truncation of 84 amino acids at the carboxyl terminus in proteins ma1-G41 and ma1-GD compared with proteins Ma1-G41 and Ma1-GD

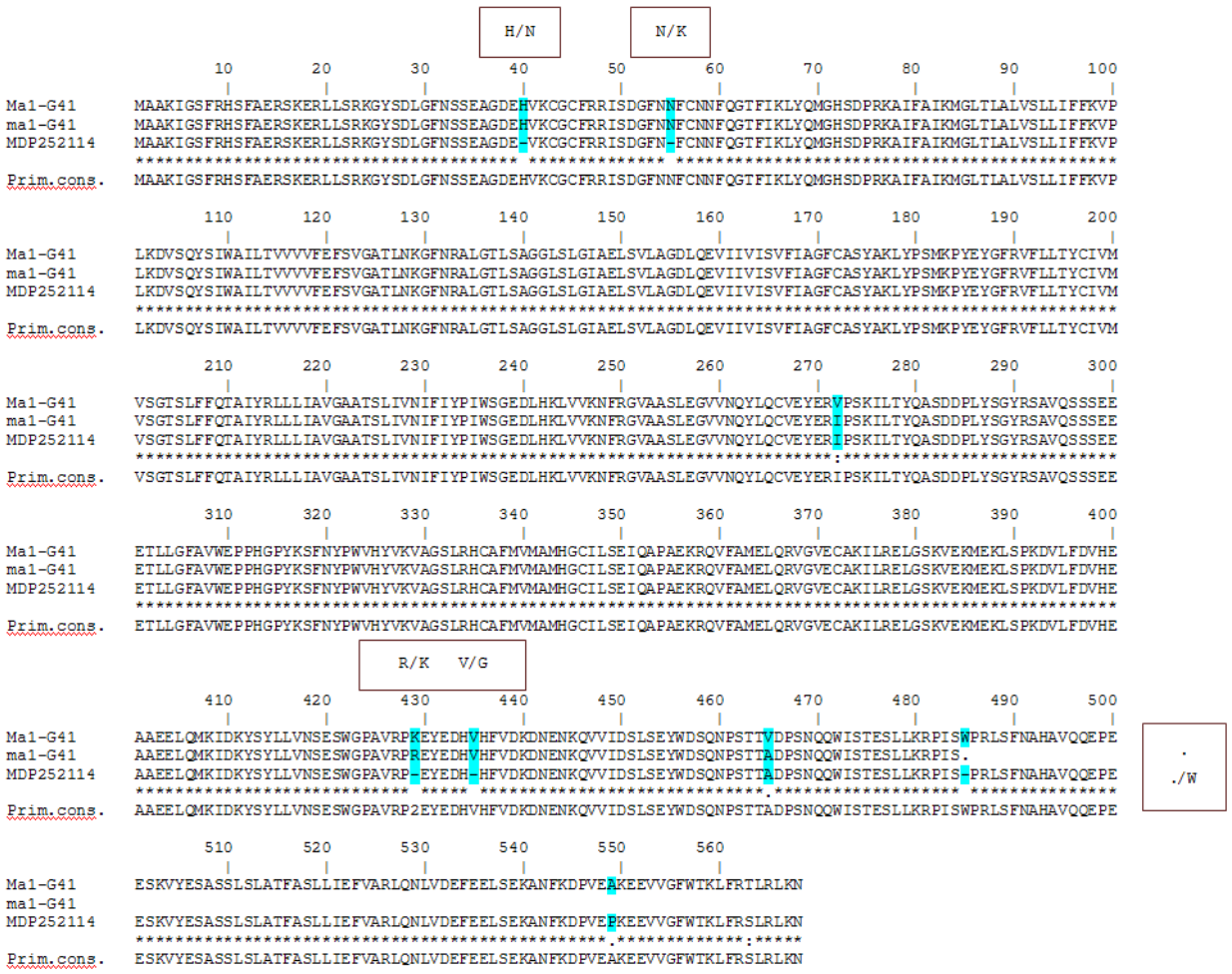

H/N
N/K
R/K
V/G
Ma1-G41
ma1-G41
MDP252114
Prim.cons.
MAAKIGSFRHSFAERSKERLLSRKGYSDLGFNSSEAGDEHVKCGCFRRISDGFNNFCNNFQGTFIKLYQMGHSDPRKAIFAIKMGLTLALVSLLIFFKVP
MAAKIGSFRHSFAERSKERLLSRKGYSDLGFNSSEAGDEHVKCGCFRRISDGFNNFCNNFQGTFIKLYQMGHSDPRKAIFAIKMGLTLALVSLLIFFKVP
MAAKIGSFRHSFAERSKERLLSRKGYSDLGFNSSEAGDE-VKCGCFRRISDGFN-FCNNFQGTFIKLYQMGHSDPRKAIFAIKMGLTLALVSLLIFFKVP
MAAKIGSFRHSFAERSKERLLSRKGYSDLGFNSSEAGDEHVKCGCFRRISDGFNNFCNNFQGTFIKLYQMGHSDPRKAIFAIKMGLTLALVSLLIFFKVP
LKDVSQYSIWAILTVVVVFEFSVGATLNKGFNRALGTLSAGGLSLGIAELSVLAGDLQEVIIVISVFIAGFCASYAKLYPSMKPYEYGFRVFLLTYCIVM
LKDVSQYSIWAILTVVVVFEFSVGATLNKGFNRALGTLSAGGLSLGIAELSVLAGDLQEVIIVISVFIAGFCASYAKLYPSMKPYEYGFRVFLLTYCIVM
LKDVSQYSIWAILTVVVVFEFSVGATLNKGFNRALGTLSAGGLSLGIAELSVLAGDLQEVIIVISVFIAGFCASYAKLYPSMKPYEYGFRVFLLTYCIVM
LKDVSQYSIWAILTVVVVFEFSVGATLNKGFNRALGTLSAGGLSLGIAELSVLAGDLQEVIIVISVFIAGFCASYAKLYPSMKPYEYGFRVFLLTYCIVM
VSGTSLFFQTAIYRLLLIAVGAATSLIVNIFIYPIWSGEDLHKLVVKNFRGVAASLEGVVNQYLQCVEYERVPSKILTYQASDDPLYSGYRSAVQSSSEE
VSGTSLFFQTAIYRLLLIAVGAATSLIVNIFIYPIWSGEDLHKLVVKNFRGVAASLEGVVNQYLQCVEYERIPSKILTYQASDDPLYSGYRSAVQSSSEE
VSGTSLFFQTAIYRLLLIAVGAATSLIVNIFIYPIWSGEDLHKLVVKNFRGVAASLEGVVNQYLQCVEYERIPSKILTYQASDDPLYSGYRSAVQSSSEE
VSGTSLFFQTAIYRLLLIAVGAATSLIVNIFIYPIWSGEDLHKLVVKNFRGVAASLEGVVNQYLQCVEYERIPSKILTYQASDDPLYSGYRSAVQSSSEE
ETLLGFAVWEPPHGPYKSFNYPWVHYVKVAGSLRHCAFMVMAMHGCILSEIQAPAEKRQVFAMELQRVGVECAKILRELGSKVEKMEKLSPKDVLFDVHE
ETLLGFAVWEPPHGPYKSFNYPWVHYVKVAGSLRHCAFMVMAMHGCILSEIQAPAEKRQVFAMELQRVGVECAKILRELGSKVEKMEKLSPKDVLFDVHE
ETLLGFAVWEPPHGPYKSFNYPWVHYVKVAGSLRHCAFMVMAMHGCILSEIQAPAEKRQVFAMELQRVGVECAKILRELGSKVEKMEKLSPKDVLFDVHE
ETLLGFAVWEPPHGPYKSFNYPWVHYVKVAGSLRHCAFMVMAMHGCILSEIQAPAEKRQVFAMELQRVGVECAKILRELGSKVEKMEKLSPKDVLFDVHE
AAEELQMKIDKYSYLLVNSESWGPAVRPKEYEDHVHFVDKDNENKQVVIDSLSEYWDSQNPSTTVDPSNQQWISTESLLKRPISNPRLSFNAHAVQQEPE
AAEELQMKIDKYSYLLVNSESWGPAVRPREYEDHVHFVDKDNENKQVVIDSLSEYWDSQNPSTTADPSNQQWISTESLLKRPIS.
AAEELQMKIDKYSYLLVNSESWGPAVRP-EYEDH-HFVDKDNENKQVVIDSLSEYWDSQNPSTTADPSNQQWISTESLLKRPIS-PRLSFNAHAVQQEPE
AAEELQMKIDKYSYLLVNSESWGPAVRP2EYEDHVHFVDKDNENKQVVIDSLSEYWDSQNPSTTADPSNQQWISTESLLKRPISWPRLSFNAHAVQQEPE
ESKVYESASSLSLATFASLLIEFVARLQNLVDEFEELSEKANFKDPVEAKEEVVGFWTKLFRTLRLKN
ESKVYESASSLSLATFASLLIEFVARLQNLVDEFEELSEKANFKDPVEPKEEVVGFWTKLFRSLRLKN
ESKVYESASSLSLATFASLLIEFVARLQNLVDEFEELSEKANFKDPVEAKEEVVGFWTKLFRSLRLKN
.
./W

Figure 1.8 Alignment of the Ma2 deduced protein sequences. MDP244249 stands for the Golden Delicious protein MDP0000244249, which combines proteins Ma2-GD and ma2-GD. Each sign "-" in MDP244249 is for two possible amino acid residues, and annotated accordingly as shown. Amino acid residues that vary are highlighted in blue

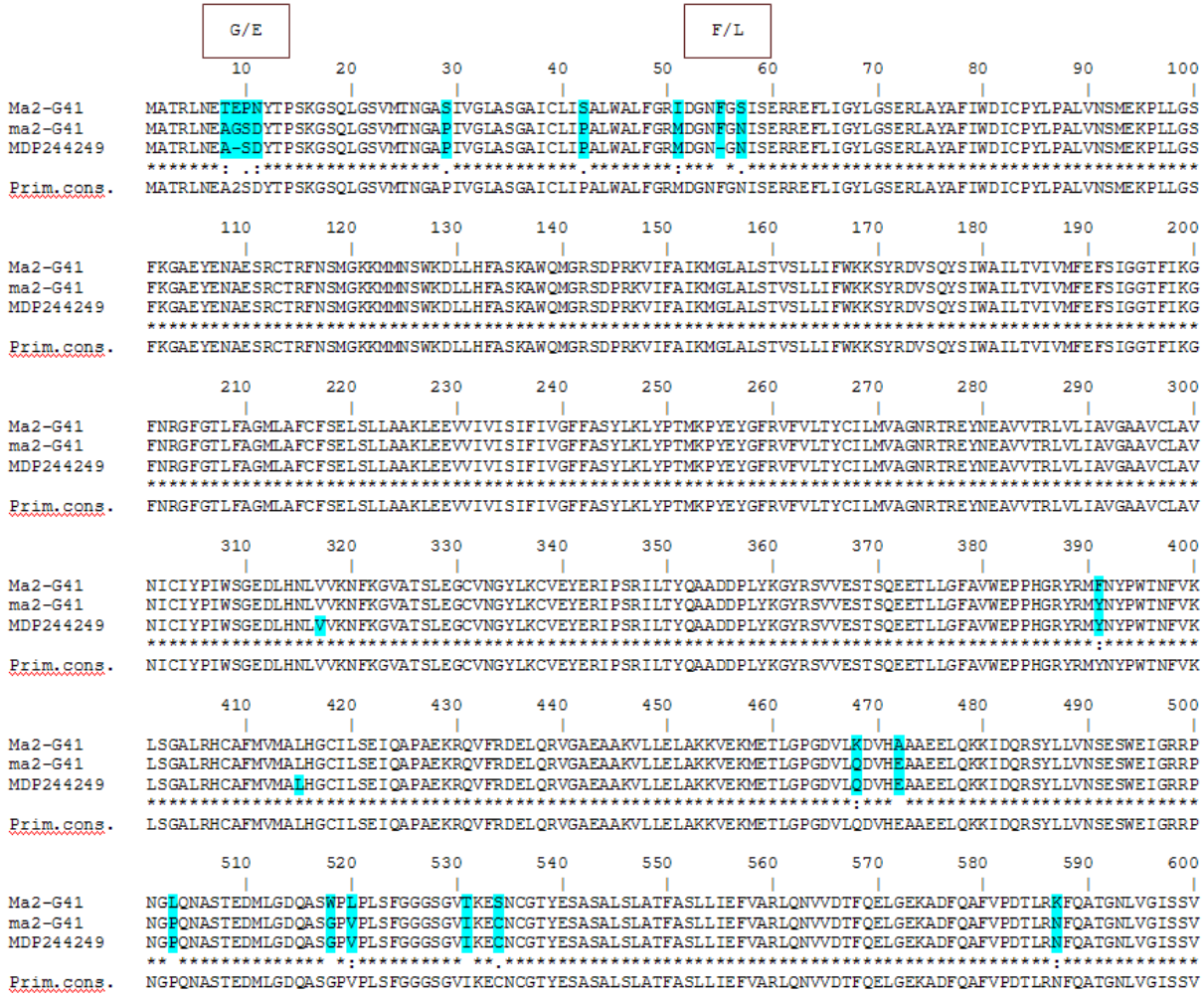

G/E
F/L
Ma2-G41
ma2-G41
MDP244249
Prim.cons.

Table 1.8 DNA and amino acid sequence variations in the *Ma2* alleles of G.41[a] and Golden Delicious (GD)[b]

Base position[c]	DNA					AA				
	Ma2-G41	*ma2-G41*	*Ma2-GD*	*ma2-GD*	MDP0000244249	*Ma2-G41*	*ma2-G41*	*Ma2-GD*	*ma2-GD*	MDP0000244249
22	A	G	G	G	G	T	A	A	A	A
26	A	G	A	G	R=A,G	E	G	E	G	E,G
28	C	T	T	T	T	P	S	S	S	S
31	A	G	G	G	G	N	D	D	D	D
85	T	C	C	C	C	S	P	P	P	P
124	T	C	C	C	C	S	P	P	P	P
153	T	G	G	G	G	I	M	M	M	M
165	T	T	A	T	W=A,T	F	F	L	F	F,L
170	G	A	A	A	A	S	N	N	N	N
627	C	T	T	T	T	F	F	F	F	F
885	T	C	C	C	C	A	A	A	A	A
951	A	A	T	A	W=A,T	V	V	V	V	V
1172	T	A	A	A	A	F	Y	Y	Y	Y

Table 1.8 (Continued)

1245	A	C	A	C	M=A,C	L	L	L	L	L
1251	A	G	G	G	G	G	G	G	G	G
1402	A	C	C	C	C	K	Q	Q	Q	Q
1415	C	A	A	A	A	A	E	E	E	E
1508	T	C	C	C	C	L	P	P	P	P
1552	T	G	G	G	G	W	G	G	G	G
1557	A	C	C	C	C	P	P	P	P	P
1558	T	G	G	G	G	L	V	V	V	V
1592	C	T	T	T	T	T	I	I	I	I
1600	A	T	T	T	T	S	T	T	T	T
1761	A	T	T	T	T	K	N	N	N	N

[a] There are 24 base variations in the coding sequence between *Ma2-G41* and *ma2-G41*. Out of the 24 base variations, seven are silent mutations and 17 are pronounced.

[b] There are four ambiguous nucleotides in the coding sequence of *MDP0000244249*, two of which are silent.

[c] Counting from the 1st base in the coding sequences

Figure 1.9 Agarose gel analysis of marker $CAPS_{1455}$. Bands of 2013 bp correspond to allele *$CAPS_{1455G}$*, i.e. the *Ma1-1455G* allele for high acidity. The combined bands of 1764 bp and 249 bp are expected for allele *$CAPS_{1455A}$*, i.e. the *ma1-1455A* allele for low acidity. A/A= genotype *$CAPS_{1455A}$ $CAPS_{1455A}$*, A/G= genotype *$CAPS_{1455A}$ $CAPS_{1455G}$*, and G/G= genotype *$CAPS_{1455G}$ $CAPS_{1455G}$*. The numbers indicate fruit pH. **a** Lane 1: 1 kb Plus DNA Ladder (Invitrogen, CA). Lanes 2-12: 11 progeny of GMAL 4595. **b** Lane 13: 1 kb Plus DNA Ladder. Lanes 14-22: Nine apple germplasm accessions as shown

a
1. M
2. 6-62
3. 6-64
4. 6-65
5. 6-67
6. 6-68
7. 6-69
8. 6-70
9. 6-71
10. 6-73
11. 6-74
12. 6-75
4.48 A/A
3.31 A/G
2.97 G/G
3.00 A/G
3.29 A/G
3.27 G/G
3.28 A/G
3.40 A/G
3.31 G/G
3.24 A/G
3.16 A/G
2.00Kb
1.65Kb
1.00Kb
0.30Kb
b
13. M
14. Britegold
15. Co-op 15
16. Cox's Orange Pippin
17. E7-47
18. Ein Shemer
19. Golden Delicious
20. Jonathan
21. Marshall McIntosh
22. Novosibirski Sweet
4.51 A/A
3.16 G/G
3.14 A/G
3.67 A/G
3.23 A/G
3.37 A/G
3.25 A/G
3.09 A/G
4.33 A/A

(*ma1-1455A*), such as *ma1-G41* or *ma1-GD*, with the low acid allele *ma*. Consequently, alleles *Ma2-G41* and *Ma2-G41* are associated with *Ma* while *ma2-G41* and *ma2-GD* with *ma*.

Together with the analyses in the haplotypes at the *Ma* locus and allelic variations in the two genes *Ma1* and *Ma2* in G.41 and Golden Delicious, the allelic associations identified here conclude that BAC3 stands for a haplotype of *Ma* for high acidity while BAC21 represents a haplotype of *ma* for low acidity, and that the allele diversity is higher for the high acidity alleles, but none or low for the low acidity alleles.

1.47 Association of the mutation-led truncation in *Ma1* with low fruit acidity in apple germplasm

To see how SNP_{1455} may explain the acidity levels in the other apple germplasm, a set of 29 (Table 1.2) representative apple germplasm accessions were analyzed with marker $CAPS_{1455}$ (Figures 1.9b, 1.10). Genotype $CAPS_{1455A}CAPS_{1455A}$ is associated either exclusively with high pH (7/7, Figure 1.10a) or tightly with low TA (7/9, Figure 1.10b). Genotypes $CAPS_{1455G}CAPS_{1455G}$ and $CAPS_{1455G}CAPS_{1455A}$, however, are associated either completely with low pH (22/22, Figure 1.10a) or highly with high TA (20/22, Figure 1.10b). These data indicate a complete or highly tight association between the mutation-led truncation in *Ma1* (*ma1-1455A*) with low acidity in these apple accessions.

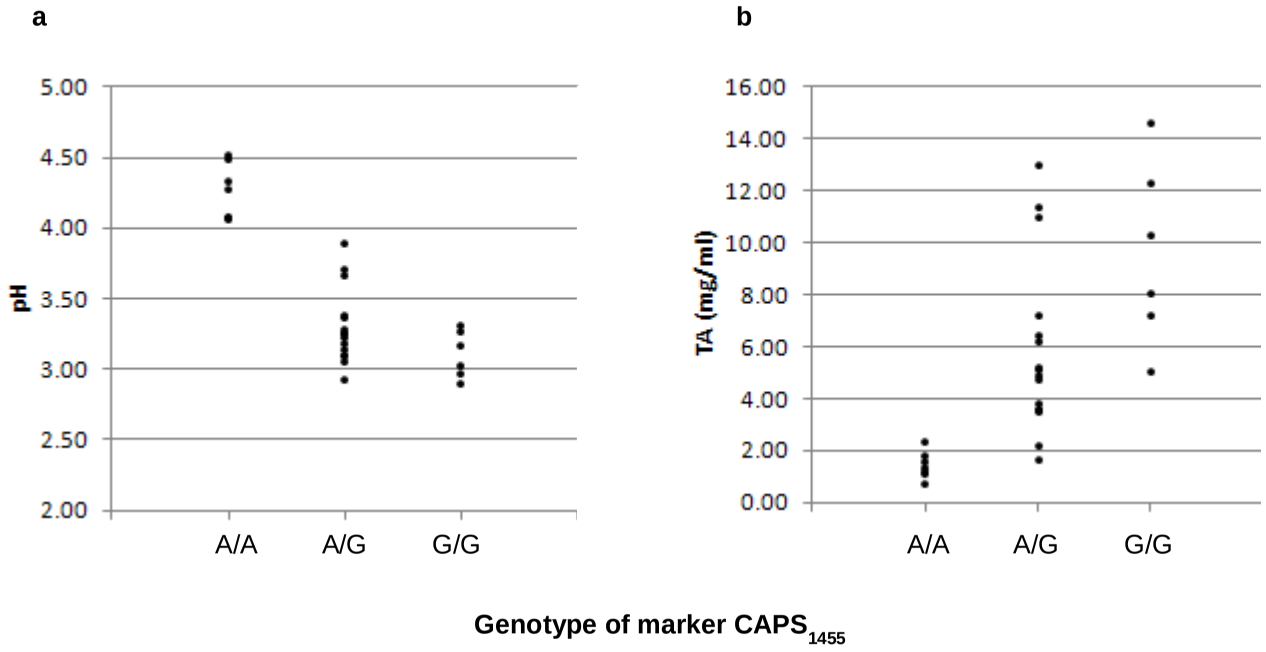


Figure 1.10 Survey of marker $CAPS_{1455}$ genotypes and their association with fruit pH **(a)** and TA (**b**) in 29 apple germplasm accessions (Table 1.2). A/A= genotype $CAPS_{1455A}$ $CAPS_{1455A}$, A/G= genotype $CAPS_{1455A}$ $CAPS_{1455G}$, and G/G= genotype $CAPS_{1455G}$ $CAPS_{1455G}$

1.5 Discussion

1.5.1 Delimiting the *Ma* locus to a 65 kb genomic segment and identification of two ALMT-like genes *Ma1* and *Ma2*

By developing three new markers and analyzing two additional populations, we delimited the *Ma* locus between markers CN889255SNP and 12514.266. The genetic interval was supported by four recombinants with GMAL 4595-6-149 and GMAL 4590-1-131 between marker CN889255SNP and *Ma*, and GMAL 4592-4-33 and GMAL 4595-6-121 between *Ma* and marker 12514.266 (Figures 1.1a, 1.2) among the 52 informative recombinants identified. The *Ma* locus between markers CN889255SNP and 12514.266 corresponds to a homologous genomic segment of 65 kb in Golden Delicious, enabling us to reduce the number of candidate genes of *Ma* from 44 identified previously (Xu et al 2011) to 19 in the present study.

Since the draft sequence of the apple genome does not provide clear haplotype specific information (Velasco et al. 2010), we identified two BAC clones of different haploid origin from apple rootstock G.41, BAC3 and BAC21, which completely cover the *Ma* locus. Sequencing the two BAC clones revealed that the *Ma* locus spanned 71 kb in BAC3 and 82 kb in BAC21. A more detailed analysis showed that out of the 19 predicted genes in Golden Delicious, three were not present in the two BACs and two were beyond the *Ma* interval. In the remaining 14 predicted genes, two were duplicated, leading to 12 predicted genes for *Ma* in both BACs, including *Ma1* and *Ma2* (Figure 1.1g-I; Table 1.5). Although the draft sequence of the apple genome is of high quality (Velasco et al. 2010), the local general structure of the *Ma* locus revealed by the two sequenced BACs from G.41 may be more representative. Given the limited number of genes in the *Ma* locus and the putative functions of ALMT genes in

maintaining the malate homeostasis in plant cells, e.g. *AtALMT9* (Kovermann et al. 2007) and *AtALMT6* (Meyer et al. 2011), *Ma1* and *Ma2* are considered to be strong candidate genes of *Ma*.

1.5.2 Putative function of Ma1 and Ma2 as vacuolar malate channels/transporters in apple fruit

The first member of the ALMT1 family unique to plants is TaALMT1 that confers wheat tolerance to soil aluminum toxicity (Sasaki et al. 2004). TaALMT1 protein facilitates malate efflux from root apices and is localized on the plasma membrane (Yamaguchi et al. 2005). The counterpart of *TaALMT1* that shows similar aluminum tolerance function includes *AtALMT1* in *Arabidopsis* (Hoekenga et al. 2006), *ScALMT1-M39.1* and *ScALMT1-1135.1* (a hybrid gene) in rye (Collins et al. 2008), and *BnALMT1* and *BnALMT2* in rape (Ligaba et al. 2006). The *Arabidopsis* genome encodes 14 *ALMT1* genes, which are distributed in four of the five clades in the ALMT1 family (Barbier-Brygoo et al. 2011). Phylogenetic analysis of the deduced protein sequences of Ma1 and Ma2 together with the 14 Arabidopsis ALMT1 proteins showed that the two apple proteins belong to clade 2 that includes five members AtALMT3-6,9 (Figure 1.11).

AtALMT9 is a vacuolar membrane protein functioning as a vacuolar malate channel for maintaining cell malate homeostasis (Kovermann et al. 2007), differing from AtALMT1 (Yamaguchi et al. 2005) and AtALMT12 (Meyer et al. 2010), which are plasma membrane proteins. AtALMT9 is expressed in all organs, but its expression in leaves is specifically in mesophyll cells. AtALMT6, another member in clade 2 that has been characterized recently, is expressed in guard cells of leaves as well as in flower organs and stems, but not in roots (Meyer et al. 2011).

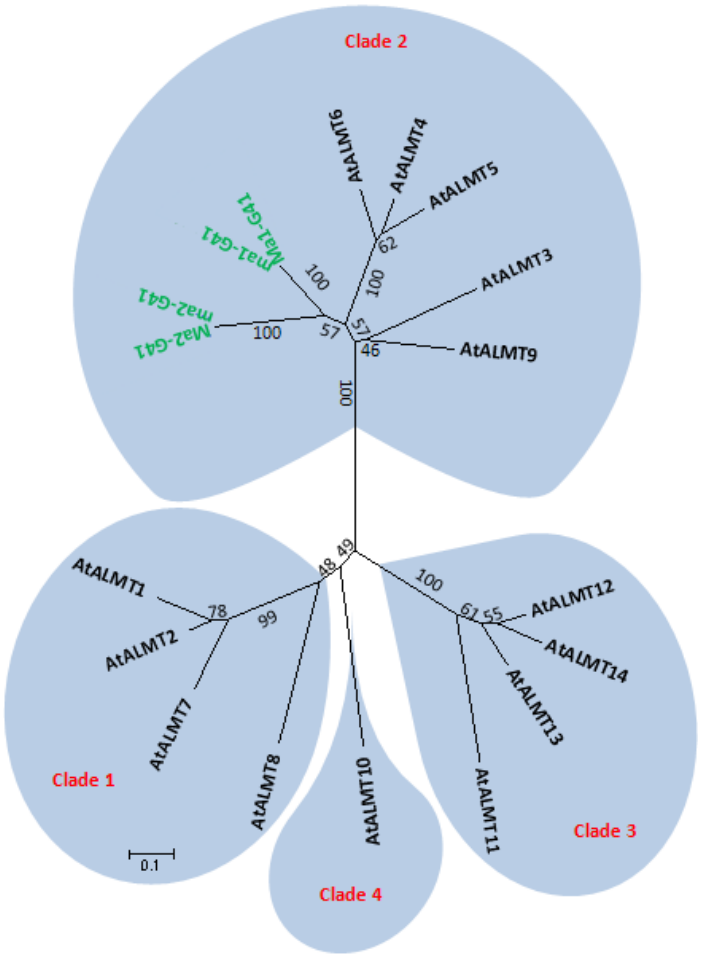


Figure 1.11 Phylogenetic analysis of Ma1 and Ma2 proteins. The 14 members AtALMT1-14 of the AtALMT1 family were retrieved from TAIR 10 (http://www.arabidopsis.org/). The protein sequences were aligned with ClustalW and the trees were constructed with the MEGA4 program (Tamura et al. 2007) using the neighbor joining method. To test the phylogeny, bootstrap samples of 1000 was set during the analysis. The tree is drawn to scale and the evolutionary distances are in the units of the number of amino acid substitutions per site. Naming system of the clades as described previously (Barbier-Brygoo et al. 2011) is adapted here

The AtALMT6 protein is also targeted to the vacuolar membrane, and it functions as a malate influx or efflux channel that is highly regulated by vacuolar pH and cytosolic malate (Meyer et al. 2011). It has been shown that low malate content in low acid fruit is the result of a restricted ability to accumulate malate in apple parenchyma cells (Beruter 2004). As members of clade 2, Ma1 and Ma2, especially Ma1, are likely vacuolar malate channels/transporters with primary function in maintaining malate homeostasis by regulating the malate levels in vacuole and cytosol in the parenchyma cells of apple fruit, thereby controlling fruit acidity levels.

1.5.3 Haplotypes of *Ma* and allelic association of the *Ma1* and *Ma2* alleles with *Ma* and *ma*

Sequencing of the two BAC clones from apple rootstock G.41 provided the first view of the *Ma* locus at the DNA sequence level with distinction between haplotypes *Ma* and *ma*. The difference between the *Ma* (BAC3) and *ma* (BAC21) haplotypes is significant in both size (82 vs. 71 kb) and the coding sequences of predicted genes (Figure 1.1g-h). In the *Ma1* and *Ma2* sequences, the alleles (*Ma1-G41* and *ma1-G41*, and *Ma2-G41* and *ma2-G41*) are clearly distinguishable. This made it possible to infer their allelic counterparts (*Ma1-GD* and *ma1-GD*, and *Ma2-GD* and *ma2-GD*) in genes *MDP0000252114* and *MDP0000244249* of Golden Delicious, respectively. Comparison of the allelic sequences of *Ma1* and *Ma2* revealed that there are no variations in the deduced amino acid sequences in alleles (*ma1-G41* and *ma1-GD*, and *ma2-G41* and *ma2-GD*) associated with *ma* for low acidity, whereas the variations are considerable for alleles (*Ma1-G41* and *Ma1-GD*, and *Ma2-G41* and *Ma2-GD*) associated with *Ma* for high acidity. A similar trend exists in the entire *Ma* region when the sequences at the *Ma* locus between Golden Delicious and G.41 were

compared as BAC21 is much closer to Golden Delicious than BAC3. One possible explanation is that the natural or human selection of fruit acidity has mostly acted upon the high acid allele *Ma* rather than the low acidity allele *ma* due to its recessive nature, leading to a greater diversity in high acidity allele *Ma*. Whether or not the high diversity among the *Ma* alleles plays a role in large fruit acidity variations in different apple cultivars would be of great interest for future investigation.

One of the most important findings of this work is the discovery of the mutation at the 1455th base of *Ma1*, which turns the tryptophan (W) codon TGG_{1455} in *Ma1-G41* into a stop codon TGA_{1455} in *ma1-G41*, leading to a premature termination and truncation of ma1-G41 by 84 deduced amino acids at the C terminus. The presence of the mutation in Golden Delicious is confirmed with the ambiguous base R_{1455}, which stands for G_{1455}/A_{1455} in *Ma1* (Figure 1.3). In view of the dramatic implication of this mutation and the critical role of the C-terminus in regulating the function and activity of TaALMT1 in wheat (Furuichi et al. 2010; Ligaba et al. 2009), marker $CAPS_{1455}$ was developed to target SNP_{1455}. Analysis using marker $CAPS_{1455}$ showed that it segregates in a codominant fashion and accurately predicts genotypes *MaMa*, *Mama* and *mama* in population GMAL 4595 and the informative recombinants (Figures 1.1, 1.2, 1.9a). Moreover, the marker shows a perfect association with pH and a highly tight association with TA in 29 apple accessions studied (Figures 1.9b, 1.10). Overall, these data strongly suggested that SNP_{1455} is critical in determining the function of the *Ma1* alleles.

It should be pointed out that the plant materials used in this study are restricted in *M. sieversii*, *M. domestica* and some of its hybrids. Since there are at least 23 species in *Malus* (Robinson et al. 2001), understanding the role of the *Ma* locus and SNP_{1455} in the remaining species would be an interesting extension of this work.

1.5.4 Expression of *Ma1* and *Ma2*

Compared with *Ma2* in expression in mature fruit, *Ma1* expression is much higher (Figure 1.6). This trend appeared to be consistent with the number of ESTs identified for the two genes in the *Malus* EST database of 336,017 accessions in GenBank. There are 20 EST accessions for *Ma1* (Table 1.6) and one for *Ma2*, i.e. CN929391 derived from pre-opened floral bud of Royal Gala. The tissue source for the 20 *Ma1* ESTs includes fruit (9 accessions), flower (3), leaf (3), root (3), stem xylem (1) and bud (1) from nine apple varieties, such as Royal Gala (6), GoldRush (5), Granny Smith (2), M.9 (1, rootstock) and others. Therefore, in addition to higher expression levels, *Ma1* is also evidenced to be expressed in a wider range of organs than *Ma2*, suggesting a broader role of *Ma1* in apple.

Significant correlations between gene expression and fruit acidity were observed for gene *Ma1* but not for *Ma2* (Figure 1.6). This suggests that *Ma1* is the major factor in determining fruit acidity levels. Since alleles *Ma1-1455G* and *ma1-1455A* are associated with *Ma* and *ma*, respectively, the strong positive correlation between *Ma1* expression and fruit acidity would suggest that transcripts of *Ma1-1455G* be more readily detected than those of *ma1-1455A*. Examining the presence of SNP_{1455} in the 20 ESTs of *Ma1* supported this reasoning. SNP_{1455G} appeared in all seven ESTs (CO723101, CX024250, CN494439, GO547092, GO509271, GO562003 and HM641023) that span over base 1455 in *Ma1* while SNP_{1455A} was not detected (Table 1.6). It appears, therefore, that both SNP_{1455} and expression levels of *Ma1* are important in apple fruit acidity. To elucidate the role of *Ma2*, more dedicated studies are needed.

Gene *MDP0000141005* encodes a putative serine/threonine protein phosphatase 2A (PP2A) regulatory subunit A and its expression was initially found to

be correlated with fruit acidity. *MDP0000141005* was excluded in allelic variation analysis since the correlation became non-significant when 18 apple accessions were analyzed. We examined the coding sequences of *MDP0000141005* in BAC3 and BAC21 of G.41, which did not confer variations in the amino acid sequences. Although PP2A is involved in many plant processes (Ahn et al. 2011; Leivar et al. 2011; Skottke et al. 2011), its subunit genes, including regulatory subunit A, have been used as reference genes for qRT-PCR analysis in plants (Czechowski et al. 2005; Navascues et al. 2012; Obrero et al. 2011). The constitutive expression of the PP2A regulatory subunit A gene and the inconsistent correlation between the *MDP0000141005* expression and fruit acidity make it unlikely the gene responsible for fruit acidity variation.

In conclusion, we discovered two ALMT-like genes, *Ma1* and *Ma2*, at the *Ma* locus of 65-82 kb containing 12-19 predicted genes that controls fruit acidity levels in apple. Expressions of *Ma1* and *Ma2* contrast sharply in the 18 apple germplasm accessions studied. *Ma1* was expressed at much higher levels than *Ma2* in mature fruit, especially in those of high acidity. Moreover, the *Ma1* expression is significantly correlated with fruit acidity, whereas the *Ma2* expression remains at low levels regardless of fruit acidity variations. These data suggest that *Ma1* is the major determinant at the *Ma* locus controlling fruit acidity. Sequencing of clones BAC3 and BAC21 that cover the two distinct haplotypes at the *Ma* locus allowed us to determine specific alleles of both *Ma1* and *Ma2* for high or low acid phenotype. A single nucleotide mutation at base 1455 in the open reading frame of *Ma1* led to a premature stop codon TGA_{1455}, which truncates the carboxyl terminus of Ma1 by 84 amino acids. A survey of 29 apple germplasm accessions using marker $CAPS_{1455}$ targeting SNP_{1455} found that the $CAPS_{1455A}$ allele is associated completely with high pH and tightly with

low TA, suggesting that the natural mutation-led truncation is most likely responsible for the abolished function of *Ma* for low pH or high TA in apple.

1.6 Reference

Ahn CS, Han JA, Lee HS, Lee S, Pai HS (2011) The PP2A regulatory subunit Tap46, a component of the TOR signaling pathway, modulates growth and metabolism in plants. Plant Cell 23:185-209

Barbier-Brygoo H, De Angeli A, Filleur S, Frachisse J-M, Gambale F, Thomine S, Wege S (2011) Anion channels/transporters in plants: from molecular bases to regulatory networks. Annu Rev Plant Biol 62:25-51

Beruter J (1998) Carbon partitioning in an apple mutant deficient in malic acid. Acta Hort 446:23-28

Beruter J (2004) Carbohydrate metabolism in two apple genotypes that differ in malate accumulation. J Plant Physiol 161:1011-1029

Blanpied GD, Silsby KJ (1992) Predicting harvest date windows for apples. Information Bulletin 221. Cornell Cooperative Extension, Cornell University, Ithaca

Boudehri K, Bendahmane A, Cardinet G, Troadec C, Moing A, Dirlewanger E (2009) Phenotypic and fine genetic characterization of the D locus controlling fruit acidity in peach. BMC Plant Biol 9:14

Brown AG, Harvey DM (1971) Nature and inheritance of sweetness and acidity in cultivated apple. Euphytica 20:68-80

Collins NC, Shirley NJ, Saeed M, Pallotta M, Gustafson JP (2008) An ALMT1 gene cluster controlling aluminum tolerance at the *Alt4* locus of rye (*Secale cereale* L.). Genetics 179:669-682

Cummins J, Aldwinckle H, Robinson T, Fazio G (2006) Apple tree rootstock named 'G.41'. In: Office TUSPaT (ed) The United States Patent and Trademark Office. Cornell Research Foundation Inc, The United States. USPP17139

Czechowski T, Stitt M, Altmann T, Udvardi MK, Scheible WR (2005) Genome-wide identification and testing of superior reference genes for transcript normalization in *Arabidopsis*. Plant Physiol 139:5-17

Emmerlich V, Linka N, Reinhold T, Hurth MA, Traub M, Martinoia E, Neuhaus HE (2003) The plant homolog to the human sodium/dicarboxylic cotransporter is the vacuolar malate carrier. P Natl Acad Sci USA 100:11122-11126

Fang DQ, Federici CT, Roose ML (1997) Development of molecular markers linked to a gene controlling fruit acidity in citrus. Genome 40:841-849

Forsline PL, Aldwinckle HS, Dickson EE, Luby JJ, Hokanson SC (2003) Collection, maintenance, characterization and utilization of wild apples of Central Asia. Horticultural Rev 29:1-61

Fulton TM, Bucheli P, Voirol E, Lopez J, Petiard V, Tanksley SD (2002) Quantitative trait loci (QTL) affecting sugars, organic acids and other biochemical properties possibly contributing to flavor, identified in four advanced backcross populations of tomato. Euphytica 127:163-177

Furuichi T, Sasaki T, Tsuchiya Y, Ryan PR, Delhaize E, Yamamoto Y (2010) An extracellular hydrophilic carboxy-terminal domain regulates the activity of TaALMT1, the aluminum-activated malate transport protein of wheat. Plant J 64:47-55

Hoekenga OA, Maron LG, Pineros MA, Cancado GMA, Shaff J, Kobayashi Y, Ryan PR, Dong B, Delhaize E, Sasaki T, Matsumoto H, Yamamoto Y, Koyama H, Kochian LV (2006) AtALMT1, which encodes a malate transporter, is identified as one of several genes critical for aluminum tolerance in *Arabidopsis*. P Natl Acad Sci USA 103:9738-9743

Hulme AC, Wooltorton LSC (1957) The organic acid metabolism of apple fruits: changes in individual acids during growth on the tree. J Sci Food Agr 8:117-122

Iwanami H, Moriya S, Kotoda N, Mimida N, Takahashi-Sumiyoshi S, Abe K (2012) Mode of inheritance in fruit acidity in apple analysed with a mixed model of a major gene and polygenes using large complex pedigree. Plant Breeding 131:322-328

Jalikop SH (2007) Linked dominant alleles or inter-locus interaction results in a major shift in pomegranate fruit acidity of 'Ganesh' × 'Kabul Yellow'. Euphytica 158:201-207

Kenis K, Keulemans J, Davey M (2008) Identification and stability of QTLs for fruit quality traits in apple. Tree Genet Genomes 4:647-661

Kovermann P, Meyer S, Hortensteiner S, Picco C, Scholz-Starke J, Ravera S, Lee Y, Martinoia E (2007) The *Arabidopsis* vacuolar malate channel is a member of the ALMT family. Plant J 52:1169-1180

Leivar P, Antolin-Llovera M, Ferrero S, Closa M, Arro M, Ferrer A, Boronat A, Camposa N (2011) Multilevel control of *Arabidopsis* 3-hydroxy-3-methylglutaryl Coenzyme A reductase by protein phosphatase 2A. Plant Cell 23:1494-1511

Lerceteau-Köhler E, Moing A, Guérin G, Renaud C, Petit A, Rothan C, Denoyes B (2012) Genetic dissection of fruit quality traits in the octoploid cultivated strawberry highlights the role of homoeo-QTL in their control. Theor Appl Genet 124:1059-1077

Liebhard R, Kellerhals M, Pfammatter W, Jertmini M, Gessler C (2003) Mapping quantitative physiological traits in apple (*Malus* × *domestica* Borkh.). Plant Mol Biol 52:511-526

Ligaba A, Katsuhara M, Ryan PR, Shibasaka M, Matsumoto H (2006) The BnALMT1 and BnALMT2 genes from rape encode aluminum-activated malate transporters that enhance the aluminum resistance of plant cells. Plant Physiol 142:1294-1303

Ligaba A, Kochian L, Pineros M (2009) Phosphorylation at S384 regulates the activity of the TaALMT1 malate transporter that underlies aluminum resistance in wheat. Plant J 60:411-423

Maliepaard C, Alston FH, van Arkel G, Brown LM, Chevreau E, Dunemann F, Evans KM, Gardiner S, Guilford P, van Heusden AW, Janse J, Laurens F, Lynn JR, Manganaris AG, den Nijs APM, Periam N, Rikkerink E, Roche P, Ryder C, Sansavini S, Schmidt H, Tartarini S, Verhaegh JJ, Vrielink-van Ginkel M, King GJ (1998) Aligning male and female linkage maps of apple (*Malus pumila* Mill.) using multi-allelic markers. Theor Appl Genet 97:60-73

Meyer S, Mumm P, Imes D, Endler A, Weder B, Al-Rasheid KAS, Geiger D, Marten I, Martinoia E, Hedrich R (2010) AtALMT12 represents an R-type anion channel required for stomatal movement in *Arabidopsis* guard cells. Plant J 63:1054-1062

Meyer S, Scholz-Starke J, De Angeli A, Kovermann P, Burla B, Gambale F, Martinoia E (2011) Malate transport by the vacuolar AtALMT6 channel in guard cells is subject to multiple regulation. Plant J 67:247-257

Navascues J, Perez-Rontome C, Sanchez DH, Staudinger C, Wienkoop S, Rellan-Alvarez R, Becana M (2012) Oxidative stress is a consequence, not a cause, of aluminum toxicity in the forage *legume Lotus corniculatus*. New Phytol 193:625-636

Nybom N (1959) On the inheritance of acidity in cultivated apples. Hereditas 45:332-350

Obrero A, Die JV, Roman B, Gomez P, Nadal S, Gonzalez-Verdejo CI (2011) Selection of reference genes for gene expression studies in Zucchini (*Cucurbita pepo*) using qPCR. J Agr Food Chem 59:5402-5411

Pfaffl MW (2001) A new mathematical model for relative quantification in real-time RT-PCR. Nucleic Acids Res 29:e45

Robinson JP, Harris SA, Juniper BE (2001) Taxonomy of the genus *Malus* Mill. (Rosaceae) with emphasis on the cultivated apple, *Malus domestica* Borkh. Plant Syst and Evol 226:35-58

Sasaki T, Yamamoto Y, Ezaki B, Katsuhara M, Ahn SJ, Ryan PR, Delhaize E, Matsumoto H (2004) A wheat gene encoding an aluminum-activated malate transporter. Plant J 37:645-653

Schumacher K, Krebs M (2010) The V-ATPase: small cargo, large effects. Curr Opin Plant Biol 13:724-730

Skottke KR, Yoon GM, Kieber JJ, DeLong A (2011) Protein phosphatase 2A controls ethylene biosynthesis by differentially regulating the turnover of ACC synthase isoforms. PLoS Genet 7: e1001370

Soglio V, Costa F, Molthoff J, Weemen-Hendriks W, Schouten H, Gianfranceschi L (2009) Transcription analysis of apple fruit development using cDNA microarrays. Tree Genet Genomes 5:685-698

Sweetman C, Deluc LG, Cramer GR, Ford CM, Soole KL (2009) Regulation of malate metabolism in grape berry and other developing fruits. Phytochemistry 70:1329-1344

Tamura K, Dudley J, Nei M, Kumar S (2007) MEGA4: Molecular evolutionary genetics analysis (MEGA) software version 4.0. Mol Biol Evol 24:1596-1599

Ulrich R (1970) Organic acids. In: Hulme A (ed) The biochemistry of fruit and their products. Academic Press, London and New York, pp 89-118

Velasco R, Zharkikh A, Affourtit J, Dhingra A, Cestaro A, Kalyanaraman A, Fontana P, Bhatnagar SK, Troggio M, Pruss D, Salvi S, Pindo M, Baldi P, Castelletti S, Cavaiuolo M, Coppola G, Costa F, Cova V, Dal Ri A, Goremykin V, Komjanc M, Longhi S, Magnago P, Malacarne G, Malnoy M, Micheletti D, Moretto M, Perazzolli M, Si-Ammour A, Vezzulli S, Zini E, Eldredge G, Fitzgerald LM, Gutin N, Lanchbury J, Macalma T, Mitchell JT, Reid J, Wardell B, Kodira C, Chen Z, Desany B, Niazi F, Palmer M, Koepke T, Jiwan D, Schaeffer S, Krishnan V, Wu C, Chu VT, King ST, Vick J, Tao Q, Mraz A, Stormo A, Stormo K, Bogden R, Ederle D, Stella A, Vecchietti A, Kater MM, Masiero S, Lasserre P, Lespinasse Y, Allan AC, Bus V, Chagne D, Crowhurst RN, Gleave AP, Lavezzo E, Fawcett JA, Proost S, Rouze P, Sterck L, Toppo S, Lazzari B, Hellens RP, Durel C-E, Gutin A, Bumgarner RE, Gardiner SE, Skolnick M, Egholm M, Van de Peer Y, Salamini F, Viola R (2010) The genome of the domesticated apple (*Malus* × *domestica* Borkh.). Nat Genet 42:833-839

Visser T, Verhaegh JJ (1978) Inheritance and selection of some fruit characters of apple .1. Inheritance of low and high acidity. Euphytica 27:753-760

Xu K, Wang A, Brown S (2011) Genetic characterization of the *Ma* locus with pH and titratable acidity in apple. Mol Breeding DOI: 10.1007/s11032-011-9674-7

Yamaguchi M, Sasaki T, Sivaguru M, Yamamoto Y, Osawa H, Ahn SJ, Matsumoto H (2005) Evidence for the plasma membrane localization of Al-activated malate transporter (ALMT1). Plant Cell Physiol 46:812-816

Yao Y-X, Li M, Zhai H, You C-X, Hao Y-J (2011) Isolation and characterization of an apple cytosolic malate dehydrogenase gene reveal its function in malate synthesis. J Plant Physiol 168:474-480

Yao YX, Li M, Liu Z, Hao YJ, Zhai H (2007) A novel gene, screened by cDNA-AFLP approach, contributes to lowering the acidity of fruit in apple. Plant Physiol Bioch 45:139-145

Yao YX, Li M, Liu Z, You CX, Wang DM, Zhai H, Hao YJ (2009) Molecular cloning of three malic acid related genes MdPEPC, MdVHA-A, MdcyME and their expression analysis in apple fruits. Sci Hortic 122:404-408

You F, Huo N, Gu Y, Luo M-c, Ma Y, Hane D, Lazo G, Dvorak J, Anderson O (2008) BatchPrimer3: A high throughput web application for PCR and sequencing primer design. BMC Bioinformatics 9:253

Zhang YZ, Li PM, Cheng LL (2010) Developmental changes of carbohydrates, organic acids, amino acids, and phenolic compounds in 'Honeycrisp' apple flesh. Food Chem 123:1013-1018

CHAPTER 2

TOWARDS AN IMPROVED APPLE REFERENCE TRANSCRIPTOME USING RNA-SEQ[1]

2.1 Abstract

The reference genome of apple (*Malus* × *domestica*) has been available since 2010. Despite being a milestone in apple genomics, the reference genome is difficult to be used as a reference in RNA-seq (RNA sequencing) analysis, a widespread technology in transcriptomics studies. One of the major limitations appears to be the low coverage of the reference transcriptome in RNA-seq mapping of reads. To improve the reference transcriptome, we obtained 14 sets of strand specific RNA-seq data of 168.5 million reads in total from fruit of Golden Delicious (GD, the source of the reference genome) in varying growth and developmental stages. Using a combination of genome-guided assembly and de novo assembly, the apple reference transcriptome was improved to a collection of 71,178 genes or transcripts, which includes 53,654 genes predicted originally (with MDP prefixed in their IDs) and 17,524 novel transcripts. Of these novel transcripts, 8,144 were identified from reads directly mapped to the reference genome while the remaining 9,380 were extracted from de novo assemblies of reads that could not be initially mapped to the reference genome.

[1] Yang Bai, Laura Dougherty, Kenong Xu (2014). Towards an improved apple reference transcriptome using RNA-seq Molecular Genetics and Genomics DOI 10.1007/s00438-014-0819-3

Evaluating the improved apple reference transcriptome with reads from Golden Delicious and other genotypes used in this and other studies showed that it allowed 62.5±9.3%-82.3±2.7% of reads to be mapped, a marked increase from the low rates of 37.4±7.7%-46.6±7.1% offered by the original reference transcriptome. The improved reference transcriptome therefore represents a step forward towards a complete reference transcriptome in apple.

2.2 Introduction

The development of RNA sequencing (RNA-seq) technology (Mortazavi et al. 2008; Wilhelm and Landry 2009) has been a breakthrough in the characterization of complex eukaryotic transcriptomes. Compared with micro-array based global gene expression assays, which primarily employ molecular hybridization between a sample of unknown transcripts and an arrayed transcriptome, RNA-seq directly sequences the mRNA derived cDNA using a next generation sequencing (NGS) platform, such as Illumina HiSeq 2000/2500. The massive throughput of NGS machines allows RNA-seq to provide unprecedented resolution and depth of data, enabling simultaneous quantification of gene expression, discovery of novel transcripts and exons, detection of SNP and measurement of splicing variants (Chepelev et al. 2009; Wilhelm et al. 2010). Due to these advantages, RNA-seq has quickly become the choice in diverse transcriptomics studies attempting to investigate complex biological process in human, animal and plants as well as in microbes.

Early RNA-seq based transcriptomics studies in plants included *Arabidopsis* (Lister et al. 2008), grape (Zenoni et al. 2010), maize (Li et al. 2010) and rice (Zhang et al. 2010). It has now been used in plant species not only with a reference genome, but also without a reference genome (Ong et al. 2012; Ruttink et al. 2013). An

essential step in RNA-seq data analysis is to map the short reads back to the reference genome or reference transcriptome so that the reads associated with a specific gene could be counted and then used to compare with other genes for differentiating their expression levels. When a reference genome is not available, an alternative reference transcriptome can be assembled by de novo assembly of RNA-seq reads directly (Ong et al. 2012; Ruttink et al. 2013). This alternative approach allows a much broader range of RNA-seq applications. However, chromosomal locality information and local and global genomic contexts would not be available in this alternative approach, and alternative splicing is also less readily detectable due to the nature of matured mRNA, which is the source of RNA-seq reads in most cases. In addition to the utility in identifying genes of splicing variants, nucleotide polymorphisms, and expression levels that are co-elevated or -suppressed in certain pathways, RNA-seq has also become a tool of discovery for revealing novel dimensions hidden in plant transcriptomes, such as RNA editing in chloroplast and mitochondria (Sun et al. 2013; Suzuki et al. 2013) and long non-coding RNAs that function as endogenous microRNA (miRNA) target mimics preventing miRNAs from reaching their target genes (Wu et al. 2013).

Apple (*Malus* × *domestica*, 2n=2x=34 usually) is one of the most important fruit crops in the world. Its genome sequences of 742.3 Mb (Velasco et al. 2010) are available from the Genome Databases for Rosaceae (GDR, http://www.rosaceae.org) and other sites. There are 63,541 predicted genes (or MDPs due to prefix MDP in gene IDs, e.g. MDP0000252114) in the consensus gene set in the genome. The source of the reference genome is Golden Delicious (GD), an apple variety grown widely throughout the major production areas in the US and abroad. RNA-seq based transcriptomics studies have also been reported in apple (Krost et al. 2012 and 2013; Zhang et al. 2012; Gapper et al. 2013; Gusberti et al . 2013). However, using the

predicted genes as a reference transcriptome has led to inconsistent RNA-seq reads mapping rates from 35.8±3.7% (unique reads, Gusberti et al . 2013) to 65% (Gapper et al. 2013), leaving more than one third of reads uncounted even if the non-specific reads were counted in Gusberti et al. (2013). In other RNA-seq based studies, the reference transcriptome was not used (Krost et al. 2012 and 2013; Zhang et al. 2012). Although the gene prediction in the apple genome might not be perfect, the 63,541 predicted genes that represent the current version of apple reference transcriptome are invaluable resource and have been used in many studies since the genome sequences became available in 2010. Clearly, there is a need for improving the reference transcriptome by building on it in the apple research community. To address this need, we obtained 14 sets of strand specific RNA-seq data from GD fruit in varying growth and developmental stages that were used in one of our previous studies (Wang and Xu 2012). Using a combination of genome-guided assembly and de novo assembly, the apple reference transcriptome was improved to a collection of 71,178 genes or transcripts, including 53,654 MDPs and 17,524 novel transcripts. Testing of RNA-seq mapping with reads from this and other studies indicated that the improved apple reference transcriptome increased the reads mapping rates to 62.5±9.3%-82.3±2.7% using high stringent mapping parameters, a considerable lift from the rates of 37.4±7.7%-46.6±7.1% offered by the original reference transcriptome.

2.3 Materials and Methods

2.3.1 Plant materials and RNA isolation

Fruit of Golden Delicious (GD) were sampled from 14 time points from 1 week after full-bloom (WAF) through 20 WAF (at harvest) in 2010 as described previously

(Wang and Xu 2012). The fruit samples were flash frozen in liquid nitrogen and stored at -80^{0}C before being used. For each sample, total RNA was isolated from 2g (young fruit)-3g (mature or near mature fruit) of ground tissues pooled from at least five fruits according to Gasic et al. (2004) with modifications: Before tissue tearer homogenization, 1ml Sarkosyl of 20% (w/v) was added to 10 ml of the extraction buffer. The extracted total RNA was dissolved in EB buffer (Qiagen, Germantown, MD) supplemented by 1× Ambion RNAsecure (Invitrogen/Life Technologies, Carlsbad, CA). To activate RNAsecure, the samples were incubated at 60^{0}C (in a water bath) for 10 min and then immediately put on ice. RNA quantity and quality were evaluated by Nanodrop 1000 (Thermo Scientific, Waltham, MA) and Bioanalyzer 2100 with RNA 6000 Nano Chip (Agilent, Santa Clara, CA) as well as a 2% agarose gel (using 1/10 RNA dilutions in EB buffer with 1× Ambion RNA secure). Immediately prior to mRNA isolation, the RNA samples were treated with DNase I (amplification grade, Invitrogen) at 37°C for 30 min followed by heat inactivation at 65°C for 15 min.

2.3.2 Strand specific RNA-seq library construction and sequencing

For each sample, 5µg total RNA was used to isolate mRNA to prepare a stand specific RNA-seq library using NEBNext Poly(A) mRNA Magnetic Isolation Module and NEBNext Ultra Directional RNA Library Prep Kit for Illumina (New England Biolabs, Ipswich, MA) following the manufacturer's protocols with minor modifications. Briefly, mRNA was extracted with 15 µl of NEBNext Magnetic Oligo $d(T)_{25}$ and fragmented in NEBNext First Strand Synthesis Buffer by heating at 94°C for 10 min. First strand cDNA was reverse transcribed form the fragmented mRNA and then used as template to synthesize double stranded cDNA with dUTP replacing

dTTP. The resulting double stand cDNA was end-repaired, dA-tailed and then ligated with NEBNext Adaptor. To remove unwanted large fragments, the adaptor-ligated cDNA was selected for size using Agencourt AMPure XP beads (Bechman Coulter, Pasadena, CA) in 0.6 volumes of the ligation reaction. For optimizing the size selection, another round of size selection was performed as described in Zhong et al. (2011), where the beads in 1.4 volumes of the cDNA solution were used. Next, the selected cDNA was digested with NEBNext USER enzyme and then enriched by PCR in the following conditions: 98°C for 30s; 14 cycles of 98°C for 10s, 65°C for 30s, 72°C for 30s; 72°C for 5 min; and then hold at 4°C. The PCR enriched cDNA libraries were purified by 1.4 volumes of Agencourt AMPure XP beads and eluted in 20µl low TE buffer (10mM Tris-HCl, pH 8.0, and 0.1mM EDTA). To estimate whether or not the libraries were within the expected size range from 250 bp to 400 bp, 2µl of the purified PCR products were analyzed by 2% agarose gel electrophoresis and then visualized with ethidium bromide-staining. If the primer-dimer band (~80bp) appeared, the libraries were purified again with 1.4 volumes of the beads. The purified libraries were then quantified by Qubit 2.0 Fluorometer using the dsDNA HS Assay Kit (Invitrogen/Life Technologies, Carlsbad, CA). The 14 multiplexed libraries with 60 ng each were pooled together for single-end sequencing of 101 bases without replication in one lane of Illumina HiSeq 2000 (Illumina, San Diego, CA) at the Cornell University Biotechnology Resource Center (Ithaca, NY).

2.3.3 Reads processing and data analysis

The 14 sequence files of 180.8 million raw reads in total were generated by the Illumina pipeline in software CASAVA v1.8 in Sanger FASTQ format (available under NCBI SRA experiment number SRX392051). Only were the high quality reads

(168.5 million) that passed the chastity filter (i.e. no more than one base call in the first 25 cycles has a chastity higher than 0.6) in the pipeline used, which accounted for 93.2±1.3% of the total raw reads of 180.8 million (Table 2.1). Data analyses were performed using CLC Genomics Workbench (CLC GW) v6.5 (CLCBio, Cambridge, Massachusetts). Three files of the apple reference genome (Velasco et al. 2010) *M. domestica* v1.0 (Md-v1.0 hereafter) were downloaded from the Genome Databases for Rosaseae (GDR, http://www.rosaceae.org). The first is the genome sequence file consisting of 122,107 contigs (MDCs hereafter); the second is the coding sequence (CDS) file for the consensus gene set containing 63,541 predicted genes (MDPs hereafter); and the third is the genome annotation file for the 63,541 genes in GFF format. The genome sequence file of 122,107 MDCs and the GFF file were combined by CLC GW to reconstruct the annotated apple reference genome (Md-v1.0) locally. To enable local BLAST search of the genome sequences, the sequences of 122,107 MDCs were converted into a BLAST database using CLC GW.

For RNA-seq mapping against reference genome Md-v1.0, we used only the gene regions defined by the MDPs (i.e. without including any bases up-or down-stream of MDPs). For convenience, the set of 63,541 MDPs were collectively called Md-v1.0-RT (apple reference transcriptome v1.0) in this study. The limit for read unspecific match to Md-v1.0-RT was set to 10. To map a read, the minimum length fraction is 0.8 and the minimum similarity is 0.98 (our empirical sequence identity threshold often effective in differentiating paralog sequences in the apple genome). These two high stringent parameters were also used in large gap read mapping of sequences against Md-v1.0. In de novo assembly, the word sizes 22 (for Step 1, Figure2.1) and 23 (for Steps 2 and 8, Figure 2.1) and bubble size of 50 were chosen automatically by the CLC de novo assembler.

Figure 2.1 Flow chart of sequence analyses conducted to improve the apple reference transcriptome. Steps are indicated by the numbers in white boxes. Round 1 includes Steps 1-6, Round 2 Steps 7-23, and Round 3 Steps 24-28. Md-1.0: apple reference genome *Malus* × *domestica* v1.0. Md-1.0-RT: apple reference transcriptome associated with Md-1.0. MDPs: genes predicted in Md-1.0 and Md-1.0-RT

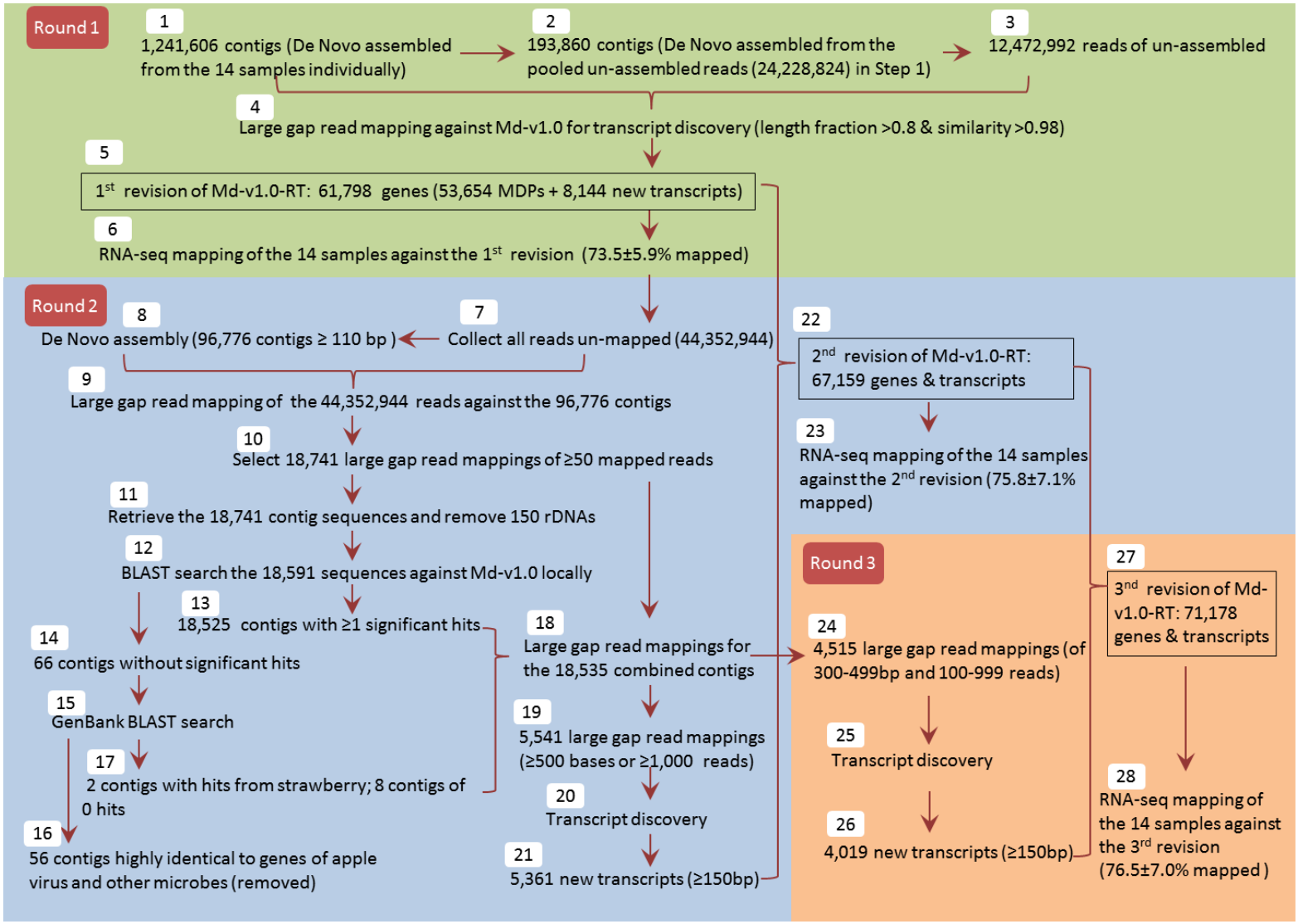

Round 1
1
1,241,606 contigs (De Novo assembled from the 14 samples individually)
2
193,860 contigs (De Novo assembled from the pooled un-assembled reads (24,228,824) in Step 1)
3
12,472,992 reads of un-assembled
4
Large gap read mapping against Md-v1.0 for transcript discovery (length fraction >0.8 & similarity >0.98)
5
1st revision of Md-v1.0-RT: 61,798 genes (53,654 MDPs + 8,144 new transcripts)
6
RNA-seq mapping of the 14 samples against the 1st revision (73.5±5.9% mapped)
Round 2
7
Collect all reads un-mapped (44,352,944)
8
De Novo assembly (96,776 contigs ≥ 110 bp)
9
Large gap read mapping of the 44,352,944 reads against the 96,776 contigs
10
Select 18,741 large gap read mappings of ≥50 mapped reads
11
Retrieve the 18,741 contig sequences and remove 150 rDNAs
12
BLAST search the 18,591 sequences against Md-v1.0 locally
13
18,525 contigs with ≥1 significant hits
14
66 contigs without significant hits
15
GenBank BLAST search
16
56 contigs highly identical to genes of apple virus and other microbes (removed)
17
2 contigs with hits from strawberry; 8 contigs of 0 hits
18
Large gap read mappings for the 18,535 combined contigs
19
5,541 large gap read mappings (≥500 bases or ≥1,000 reads)
20
Transcript discovery
21
5,361 new transcripts (≥150bp)
22
2nd revision of Md-v1.0-RT: 67,159 genes & transcripts
23
RNA-seq mapping of the 14 samples against the 2nd revision (75.8±7.1% mapped)
Round 3
24
4,515 large gap read mappings (of 300-499bp and 100-999 reads)
25
Transcript discovery
26
4,019 new transcripts (≥150bp)
27
3rd revision of Md-v1.0-RT: 71,178 genes & transcripts
28
RNA-seq mapping of the 14 samples against the 3rd revision (76.5±7.0% mapped)

Table 2.1 The number of reads in raw, filter-passed and filtered (removed)

Samples	Raw reads	Passed reads	Passed reads (%)	Filtered reads	Filtered reads (%)
WAF01	14,057,160	13,212,078	94.0	845,082	6.0
WAF02	12,306,301	11,577,420	94.1	728,881	5.9
WAF03	13,909,260	13,064,054	93.9	845,206	6.1
WAF04	13,789,342	12,902,501	93.6	886,841	6.4
WAF05	14,357,928	13,594,031	94.7	763,897	5.3
WAF06	13,952,613	13,063,572	93.6	889,041	6.4
WAF08	12,684,060	11,692,039	92.2	992,021	7.8
WAF10	12,820,062	12,036,125	93.9	783,937	6.1
WAF12	11,499,466	10,788,214	93.8	711,252	6.2
WAF14	12,313,776	11,343,438	92.1	970,338	7.9
WAF16	11,698,244	10,698,751	91.5	999,493	8.5
WAF18	13,012,946	11,701,713	89.9	1,311,233	10.1
WAF19	11,512,427	10,729,005	93.2	783,422	6.8
WAF20	12,900,512	12,124,220	94.0	776,292	6.0
Total	180,814,097	168,527,161	/	12,286,936	/
Mean	12,915,293	12,037,654	93.2	877,638	6.8
SD	979,475	986,658	1.3	156,287	1.3

For transcript discovery, parameters were mostly set by default, but with the following changes: (1) the minimum length of ORF calling was raised from 100bp to 150bp; (2) For "gene" discovery, the maximum distance between events was set to 10bp and the minimum length of "gene" was 150bp.

2.3.4 Revision of the reference genome transcriptome (Md-v1.0-RT)

An approach of three rounds of de novo assembly, large gap read mapping and/or transcript discovery was taken for improving Md-v1.0-RT (Figure 2.1). Briefly, Steps 1-5 in Round 1 were intended to reveal new transcripts directly from reference genome Md-v1.0 so that a revision of Md-v1.0-RT could be made straightforwardly. We began with a complexity reduction step by de novo assembling of the RNA-seq reads into contigs in each of the 14 samples, generating a set of contigs of 1,241,606. The unassembled reads were collected and pooled and were de novo assembled again, yielding another set of contigs of 193,860. The reads (12,472,992) that still remained unassembled were re-collected, and used along with the two sets of contig sequences (1,435,466 in sum) as input for large gap read mapping against reference genome Md-v1.0 using the CLC Large Gap Read Mapping tool. Revision of the reference transcriptome was conducted using the CLC Transcript Discovery tool, leading to the first revision of Md-v1.0-RT (Step 5). RNA-seq mapping evaluation of this initial revision of Md-v1.0-RT was completed using the reads from the 14 RNA-seq samples (Step 6). Steps 7 through 20 in Round 2 were attempted to discover new transcripts from the reads that could not be mapped in Step 6 so that more novel transcripts could be identified to achieve the second revision of Md-v1.0-RT. RNA-seq mapping evaluation of the second revision of Md-v1.0-RT was performed in Step 23 where the reads from the 14 samples were used again. The third round was largely a

supplementary step to Round 2 to identify new transcripts that were not selected early. Augmenting the second revision of MD-1.0-RT with the new transcripts in Round 3 resulted in the improved reference transcriptome Md-v1.0-RT, i.e. the third revision of Md-v1.0-RT.

2.3.5 MapMan gene ontology of new transcripts

The sequences of the new transcripts were retrieved to BLAST search multiple databases using the web-based search tool Mercator (http://mapman.gabipd.org/web/guest/mercator). The databases searched include: TAIR-Arabidopsis TAIR proteins (release 10), PPAP-SwissProt/UniProt Plant Proteins, CHLAMY-JGI Chlamy release 4 Augustus models, ORYZA-TIGR5 rice proteins, KOG: Clusters of orthologous eucaryotic genes database (KOG), CDD- conserved domain database, and IPR-interpro scan. The output file of Mercator not only contains the best hits in databases, but also assigns MapMan's gene ontology (Thimm et al. 2004) for each input sequence if possible.

2.3.6 Chromosomal locality and apple genome origin of new transcripts

Chromosomal locality or the apple genome origin of the new transcripts was either deduced from their associated home MDCs if they were identified in Round 1 (Figure 2.1), or evidenced from significant ($E<10^{-10}$) hits in BLAST searches against reference genome Md-v1.0 or GenBank that were conducted in Rounds 2 or 3 (Figure 2.1).

2.3.7 Evaluation of the improved reference genome with RNA-seq data from other sources

Two sets of published RNA-seq paired-end data (Table 2.2) generated from the Illumina platform were downloaded from NCBI Sequence Read Archive (SRA). The first set contained two samples ERR033805 and ERR033806 (Krost et al. 2012 and 2013). Both samples were collected from meristem tissues of apple breeding selections, but differed in their growth habit: ERR033805 was obtained from a standard selection while ERR033806 a columnar selection. The second set was a six-sample (ERR313216, ERR313217, ERR313224, ERR313225, ERR313226 and ERR313239) subset representing the 24 samples used in investigating apple scab ontogenic resistance (Gusberti et al. 2013). These samples were derived from mRNA of Golden Delicious leaves challenged by pathogen *Venturia inaequalis* or mock-inoculated by water. The two datasets were trimmed by three parameters (a quality limit of 0.05-a probability of error for a base called, an ambiguous limit of 2 and a minimum number of nucleotides of 15) to remove low quality reads or low quality bases in the reads prior to RNA-seq mapping (Table 2.2).

2.4 Results

2.4.1 Mapping of RNA-seq reads to the apple reference transcriptome (Md-v1.0-RT)

The current version of apple reference genome (Md-v1.0) was released in 2010 (Velasco et al. 2010). It comprises 122,107 MDCs (genomic contigs) that were annotated with 63,541 MDPs (predicted genes). To evaluate the coverage of Md-v1.0-RT, the 14 sets of RNA-seq data from fruit of Golden Delicious (GD) were mapped against the reference transcriptome. The mean input was 12,037,654±986,658 reads per sample, with 168,527,161 reads in total (Table 2.3). The mean reads mapping rate was 42.8±4.5% (34.4%-49.6%), including 32.6±3.4% mapped uniquely and 10.2±1.2%

Table 2.2 Details of the two published RNA-seq datasets used for evaluating the revised reference transcriptome

SRA Run #	Raw reads (in single reads)	Passed reads (in single reads)	Passed reads (%)	Filtered reads (in single reads)	Filtered reads (%)	Treatment	Source
ERR033805	89,647,896	83,210,433	92.8	6,437,463	7.2	Normal	Krost et al.
ERR033806	83,807,190	78,567,968	93.7	5,239,222	6.3	Normal	(2012 & 2013)
Total	173,455,086	161,778,401		11,676,685			
Mean	86,727,543	80,889,201	93.3	5,838,343	6.7		
SD	4,130,003	3,282,718	0.7	847,284	0.7		
ERR313216	66,241,188	65,375,676	98.7	865,512	1.3	*V. inaequalis-1*	Gusberti et al
ERR313217	65,620,796	65,418,747	99.7	202,049	0.3	CK-1	(2013)
ERR313224	72,855,500	72,496,504	99.5	358,996	0.5	*V. inaequalis-2*	
ERR313225	93,291,312	92,913,402	99.6	377,910	0.4	CK-3	
ERR313226	71,290,464	70,979,144	99.6	311,320	0.4	CK-2	
ERR313239	74,738,806	74,260,450	99.4	478,356	0.6	*V. inaequalis-3*	
Total	444,038,066	441,443,923		2,594,143			
Mean	74,006,344	73,573,987	99.4	432,357	0.6		
SD	10,116,936	10,159,711	0.4	230,555	0.4		

Table 2.3 RNA-seq mapping of reads against the current version of apple reference transcriptome (Md-v1.0-RT) and its revisions

Reference transcriptome	Reads mapping	Overall No. of reads	Mean No. of reads per sample	SD of mean No. of reads per sample	% of Total Reads	% of Total Reads-SD
MD-v1.0-RT	Mapped	72,398,775	5,171,341	792,594	42.8	4.5
	-uniquely	55,134,893	3,938,207	596,381	32.6	3.3
	-non-specifically	17,263,882	1,233,134	198,280	10.2	1.2
	Unmapped	96,128,386	6,866,313	623,257	57.2	4.5
	Total	168,527,161	12,037,654	986,658	100.0	0.0
1st revision	Mapped	124,174,217	8,869,587	1,175,967	73.5	5.9
	-uniquely	100,207,386	7,157,670	1,060,606	59.3	6.1
	-non-specifically	23,966,831	1,711,917	229,393	14.2	1.2
	Unmapped	44,352,944	3,168,067	652,320	26.5	5.9
	Total	168,527,161	12,037,654	986,658	100.0	0.0

Table 2.3 (Continued)

2nd revision	Mapped	127,996,932	9,142,638	1,298,164	75.8	7.1
	-uniquely	105,370,507	7,526,465	1,068,633	62.4	5.9
	-non-specifically	22,626,425	1,616,173	234,296,	13.4	1.3
	Unmapped	40,530,229	2,895,016	797,807	24.2	7.1
	Total	168,527,161	12,037,654	986,658	100.0	0.0
3rd revision	Mapped	129,277,684	9,234,120	1,298,218	76.5	7.0
	-uniquely	105,409,556	7,529,254	1,054,598	62.4	5.7
	-non-specifically	23,868,128	1,704,866	248,086	14.1	1.4
	Unmapped	39,249,477	2,803,534	783,221	23.5	7.0
	Total	168,527,161	12,037,654	986,658	100.0	0.0

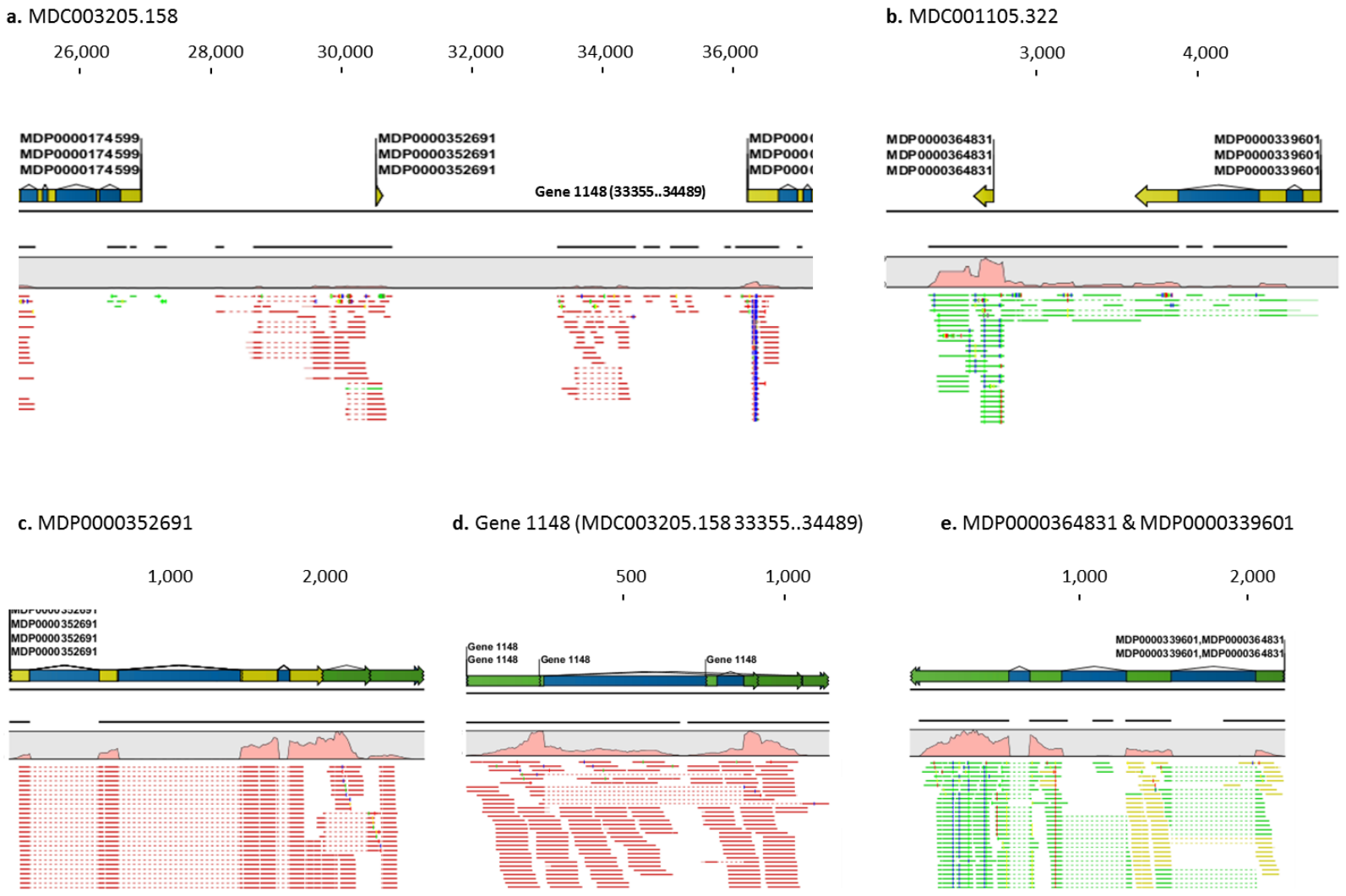
non-specifically. Of the mapped reads, 10.9±1.4% was mapped to the introns (data not shown). These data suggested that the majority (57.2±4.5%) of reads were not counted in the RNA-seq reads mapping process, and that alternative splicing variants were likely common within the MDPs as nearly 11% of mapped reads were mapped to the predicted intron regions.

2.4.2 Improvement of the apple reference transcriptome (Md-v1.0-RT)

To improve reference transcriptome Md-v1.0-RT, the 14 sets of RNA-seq data from GD fruit were used to uncover novel transcripts. An approach of three rounds of de novo assembling, large gap read mapping, transcript discovery and/or RNA-seq mapping evaluation was carried out for this purpose (Figure 2.1, Table 2.3).

In Round 1, the two sets of De Novo assembled contigs (1,435,466) and the un-assembled reads (12,472,992) from the 14 samples were mapped to 49,553 of the 122,107 MDCs in Md-v1.0 using the CLC Large Gap Read Mapping tool. The remaining 72,554 MDCs received zero reads. The CLC Transcript Discovery tool was used for the first revision of reference transcriptome Md-v1.0-RT that contained 61,798 genes or transcripts, including 53,654 MDPs and 8,144 new transcripts (Step 5, Figure 2.1) which collectively covered 176.6Mb. The 53,654 MDPs accounted for 160.7Mb (2,995±2,784bp/gene) of the transcriptome and comprised two gene sets. The first set of 53,579 genes corresponded to MDPs defined in the original Md-v1.0-RT, but with many enhanced by alternative splicing variants and/or new 5’ or 3’ sequences (Figure 2.2a,c). The second set of 74 genes that represented 154 single MDPs in the original Md-v1.0-RT as each of the 74 were combined from two or more MDPs due to the presence of bridging reads between them (Figure 2.2b,e) on the same strand. The 8,144 new transcripts had an accumulated exon length of 6.5Mb

Figure 2.2 Large gap read mapping and transcript discovery and verification. The length (bp) of MDCs and MDPs or transcripts is shown by the numbers on *top*. The *blue*, *green* and *yellow colors* in the graphic annotation of genes represent gene, mRNA and CDS (coding sequences), respectively. The mapped strand specific reads are shown in *red line* (for one strand) or *green line* (for the other stand). Reads that are not specifically mapped are indicated with *yellow lines*. The *dotted* region in reads corresponds to an intron. **a**. Large gap read mapping of reads and read contigs onto MD C003205.158 (shown a section). Note the reads and read contigs under MDP0000352691 and Gene 1148. **b**. Large gap read mapping of reads and read contigs onto MDC001105.322 (shown a section). Note the reads and read contigs that bridge genes between MDP0000364831 and MDP0000339601 on the same strand. **c**-**e**. RNA-seq mappings of reads for genes MDP0000352691 (**c**. note the new sequences expanded beyond the original coverage of MDP0000352691 in **a)**, gene 1148 (**d**. this proves gene 1148 to be a novel transcript between MDP0000352691 and MDP0000277388 in MDC003205.158) and MDP0000364831 and MDP0000339601 that were merged into one gene (**e**)


a. MDC003205.158
26,000
28,000
30,000
32,000
34,000
36,000
MDP0000174599
MDP0000174599
MDP0000174599
MDP0000352691
MDP0000352691
MDP0000352691
Gene 1148 (33355..34489)
b. MDC001105.322
3,000
4,000
MDP0000364831
MDP0000364831
MDP0000364831
MDP0000339601
MDP0000339601
MDP0000339601
c. MDP0000352691
1,000
2,000
MDP0000352691
MDP0000352691
MDP0000352691
d. Gene 1148 (MDC003205.158 33355..34489)
500
1,000
Gene 1148
Gene 1148
Gene 1148
Gene 1148
e. MDP0000364831 & MDP0000339601
1,000
2,000
MDP0000339601,MDP0000364831
MDP0000339601,MDP0000364831

(796±536bp/transcript) and represented transcribed regions that were not included in Md-v1.0-RT (Figure 2.2a,d). Alternative splicing variants were detected in the new transcripts as well (Figure 2.2d). There were 9,887 MDPs in the original Md-v1.0-RT that were excluded from in the first revision because the genomic contigs (MDCs) harboring these MDPs were among the 72,554 MDCs that received zero reads in the large gap read mapping. Repeating the mapping of the original RNA-seq data for the 14 samples using the first revision of Md-v1.0-RT resulted in mapping 73.5±5.9% of reads (Figure 2.1, Table 2.3), an increase of 30.7 percentiles from the coverage of 42.8±4.5% when using the original Md-v1.0-RT (Table 2.3).

The aim of the second round of revision was to uncover more novel transcripts. To do so, the reads that could not be mapped to the first revision of Md-v1.0-RT in RNA-seq mapping (Step 6, Figure 2.1) were collected and de novo assembled into 96,776 contigs of minimal 110 bases (Steps 7-8, Figure 2.1). When the reads were mapped back to these contigs using the CLC Large Gap Read Mapping tool, 18,741 mappings of at least 50 reads were selected (Steps 9-10, Figure 2.1). By retrieving their corresponding contig sequences and then BLAST search against local databases for rDNAs and the apple genome (Md-v1.0), 150 contigs highly similar to rDNAs were removed and 18,525 contigs of at least one significant hit ($E<10^{-10}$) in Md-v1.0 were identified (Steps 11-13, Figure 2.1). The 66 contigs without a significant hit in Mdv1.0 were BLAST searched against GenBank, leading to removing 56 of them similar to apple virus and other microbes' sequences (Steps 14-16, Figure 2.1). The remaining ten, including two with hits from strawberry and eight with zero hits, were pooled with the 18,525 contigs of hits in Md-v1.0, which allowed retrieving 18,535 corresponding large gap read mappings (Steps 13, 15, 17 and 18, Figure 2.1). Applying the CLC Transcript Discovery tool to 5,541 of the 18,535 large gap read mappings of at least 500 bases in length or at least 1,000 mapped reads revealed 5,361

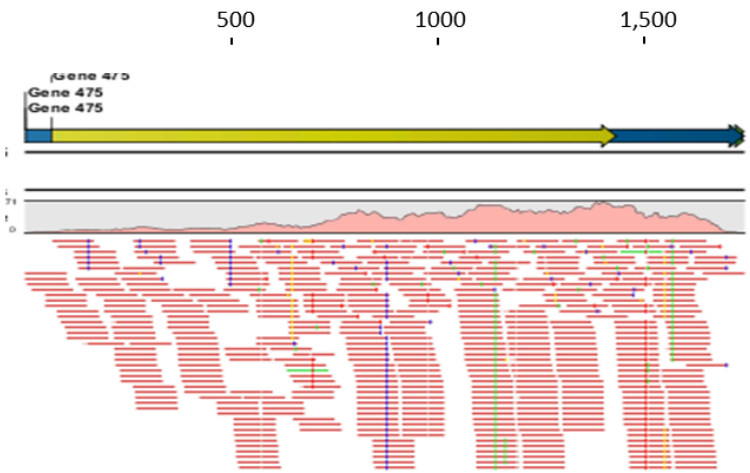
new transcripts (Steps 19-21, Figure 2.1, Figure 2.3a,b). Combining these new transcripts with the first revision of Md-v1.0-RT led to the second revision of Md-v1.0-RT (Step 22, Figure 2.1). The 5,361 new transcripts accounted for transcribed regions of 3,481kb (649±330bp/transcript) of the transcriptome and included many splice variants (Figure 2.3a,b). RNA-seq mapping of the 14 samples against the second revision of Md-v1.0-RT resulted in a read mapping rate of 75.8±7.1%, a 2.3-percentile increase over that obtained using the first revision.

The intent of the third round of revision was to identify transcripts that could account for the reads represented by the 4,515 large gap read mappings that were not selected in the second round, but had 100-999 mapped reads with contig length of 300-499 bp (Step 24, Figure 2.1). Using again the CLC Transcript Discovery tool, we obtained 4,019 new transcripts (Steps 25-26, Figure 2.1, Figure 2.3c,d) that are equivalent to an accumulated transcribed region of 1,503kb (374±58bp/transcript). Alternative splicing variants were detectable in this set of transcripts as well (Figure 2.3d). Supplementing the second revision of the reference transcriptome with these 4,019 new transcripts (Step 27, Figure 2.1) denoted the third revision of Md-v1.0-RT. Mapping of RNA-seq reads from the 14 samples to the third revision of Md-v1.0-RT resulted in a mean reads mapping rate of 76.5±7.0% (Table 2.3, Step 28, Figure 2.1). Overall, this final revision of reference transcriptome Md-v1.0-RT, which is available at the Genome Database for Rosaceae (http://www.rosaceae.org/), contained 71,178 genes or transcripts covering 172.2Mb, of which 53,654 are MDPs and 17,524 novel transcripts.

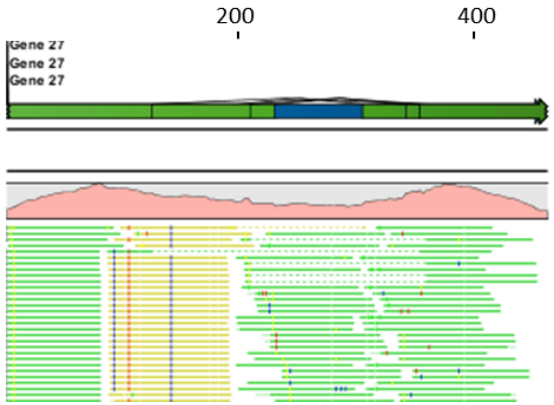
2.4.3 Evaluation of the revised apple reference transcriptome with RNA-seq data from other sources

Figure 2.3 Novel transcripts identified from reads un-mapped initially. The transcripts are annotated or presented with the same elements and colors schemes as described in Figure 2.2. **a**, **b**. Examples of new transcripts revealed in Round 2 (Figure 2.1). **c**, **d**. Examples of new transcripts uncovered in Round 3 (Figure 2.1)

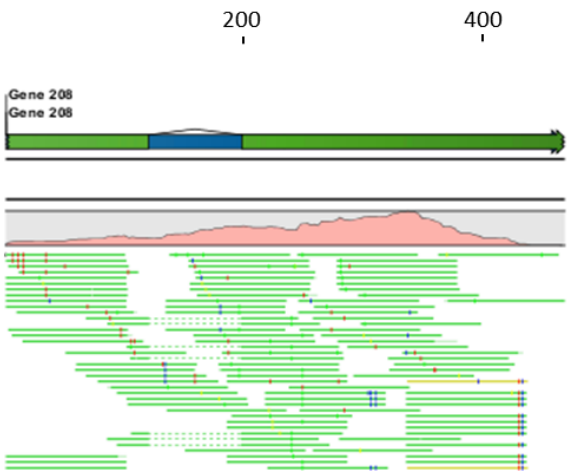
a. Gene 475



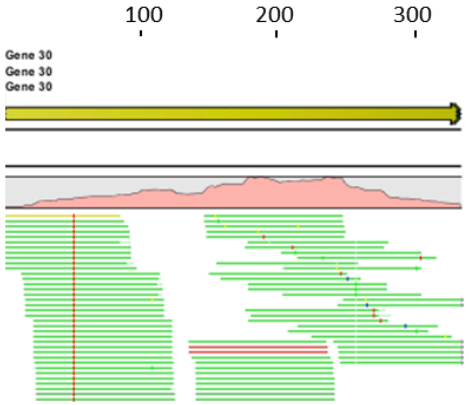
b. Gene 27



c. Gene 208



d. Gene 30



To evaluate the revised reference transcriptome, two sets of RNA-seq data that were reported in previous studies (Krost et al. 2012 and 2013; Gusberti et al 2013) were used (Table 2.2). After trimming/removing low quality bases and reads, the first dataset (ERR033805 and ERR033806) in Krost et al. (2012 and 2013) contained 80.9±3.3 million reads per sample and the second dataset (ERR313216, ERR313217, ERR313224, ERR313225, ERR313226 and ERR313239) in Gusberti et al (2013) contained 73.6±10.2 million reads per sample. RNA-seq mapping against the revised Md-v1.0-RT showed that reads of 82.3±2.7% in the first dataset and 62.6±7.7% in second dataset were mapped. In comparison, only 46.6±7.1% of reads in the first dataset and 37.4±7.7% in the second dataset were mapped using the original Md-v1.0-RT (Table 2.4). These results indicated that when reads from other tissues or genotypes were mapped with the revised reference, much higher rates of coverage were obtained.

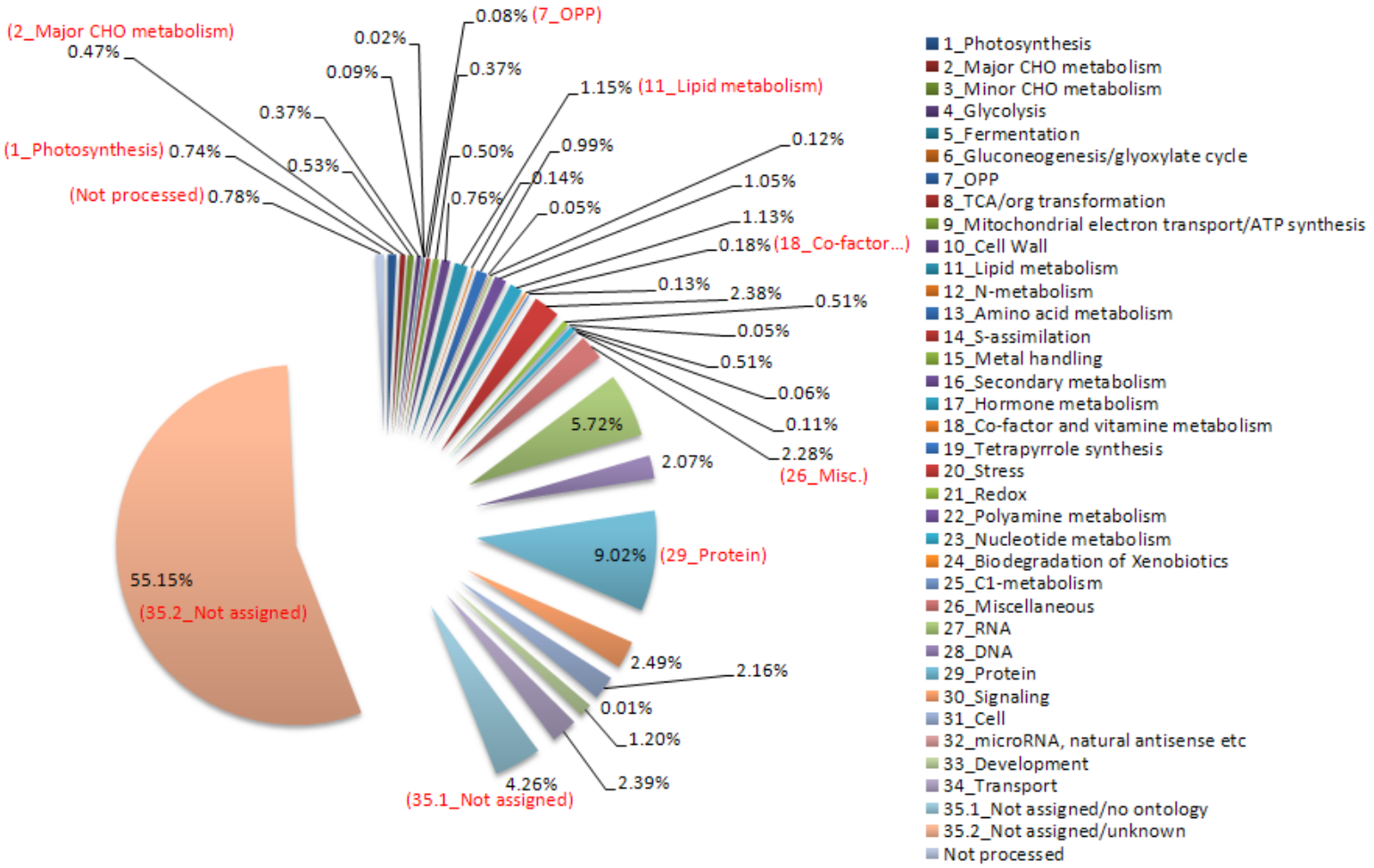
2.4.4 MapMan gene ontology of the new transcripts

Over the process of three rounds of transcript discovery, a total of 17,524 new transcripts were identified, including 8,144 identified from reads directly mapped to reference genome Md-v1.0, and 9,380 identified from the unmapped reads. To understand their putative functions as well as their possible MapMan's gene ontology (Thimm et al. 2004), multiple databases were searched with the web-based search tool Mercator (http://mapman.gabipd.org/web/guest/mercator) using the 17,524 new transcripts as BLAST queries. Of these, 7,724 (44.1%) had significant returns ($E<10^{-10}$), but only 6,978 (39.8%) were assigned with a MapMan ontology in one of the 34 bins while 746 (4.3%) were placed to 'not assigned and no ontology' (Figure 2.4). The remaining 9,800 (55.9%) were returned without significant hits or assigned

Table 2.4 Evaluation of the original and revised reference transcriptome Md-v1.0-RT with two published RNA-seq datasets

Reference transcriptome	Reads mapping	Overall (in single reads)	Mean (in single reads)	Mean-SD (in single reads)	% of Total Reads	% of Total Reads-SD	Source
Original	Mapped	74,098,570	37,049,285	7,031,597	46.6	7.1	Krost et al
	-uniquely	55,202,638	27,601,319	4,560,743	34.7	4.4	(2012 & 2013)
	-non-specifically	18,895,932	9,447,966	2,470,855	11.9	2.7	
	Unmapped	84,494,900	42,247,450	3,952,782	53.4	7.1	
	Total	158,593,470	79,296,735	3,078,815	100.0	0.0	
Revised	Mapped	130,675,496	65,337,748	4,640,958	82.3	2.7	
	-uniquely	107,283,091	53,641,546	2,432,850	67.6	0.4	
	-non-specifically	23,392,405	11,696,203	2,208,109	14.7	2.2	
	Unmapped	27,917,974	13,958,987	1,562,143	17.7	2.7	
	Total	158,593,470	79,296,735	3,078,815	100.0	0.0	
Original	Mapped	168,506,458	28,084,410	10,169,584	37.4	7.7	Gusberti et al
	-uniquely	129,605,594	21,600,932	7,812,880	28.8	5.9	(2013)
	-non-specifically	38,900,864	6,483,477	2,363,855	8.6	1.8	
	Unmapped	272,937,465	45,489,578	3,386,643	62.6	7.7	
	Total	441,443,923	73,573,987	10,159,711	100.0	0.0	
Revised	Mapped	280,188,515	46,698,086	14,131,047	62.5	9.3	
	-uniquely	238,416,607	39,736,101	11,905,174	53.2	7.8	
	-non-specifically	41,771,908	6,961,985	2,229,389	9.3	1.5	
	Unmapped	161,255,409	26,875,902	5,072,928	37.5	9.3	
	Total	441,443,923	73,573,987	10,159,711	100.0	0.0	

Figure 2.4 Distribution of the 17,524 novel transcripts in MapMan bins. The percentage was calculated from the number of transcripts in a given bin over the total transcripts of 17,524. Each bin is shown by a piece of the pie and presented clock-wide with the first and several other bins labeled in *red*. The key for each of 35 MapMan bins is listed on *right*


(2_Major CHO metabolism)
0.47%
0.02%
0.08% (7_OPP)
0.09%
0.37%
1.15% (11_Lipid metabolism)
0.37%
(1_Photosynthesis) 0.74%
0.53%
0.50%
0.99%
0.12%
(Not processed) 0.78%
0.14%
1.05%
0.76%
0.05%
1.13%
0.18% (18_Co-factor...)
0.13%
2.38%
0.51%
0.05%
0.51%
0.06%
5.72%
0.11%
2.28%
(26_Misc.)
2.07%
9.02% (29_Protein)
55.15%
(35.2_Not assigned)
2.49%
2.16%
0.01%
1.20%
4.26%
(35.1_Not assigned)
2.39%
1_Photosynthesis
2_Major CHO metabolism
3_Minor CHO metabolism
4_Glycolysis
5_Fermentation
6_Gluconeogenesis/glyoxylate cycle
7_OPP
8_TCA/org transformation
9_Mitochondrial electron transport/ATP synthesis
10_Cell Wall
11_Lipid metabolism
12_N-metabolism
13_Amino acid metabolism
14_S-assimilation
15_Metal handling
16_Secondary metabolism
17_Hormone metabolism
18_Co-factor and vitamine metabolism
19_Tetrapyrrole synthesis
20_Stress
21_Redox
22_Polyamine metabolism
23_Nucleotide metabolism
24_Biodegradation of Xenobiotics
25_C1-metabolism
26_Miscellaneous
27_RNA
28_DNA
29_Protein
30_Signaling
31_Cell
32_microRNA, natural antisense etc
33_Development
34_Transport
35.1_Not assigned/no ontology
35.2_Not assigned/unknown
Not processed

to category ‘not assigned and unknown’ (9,664 or 55.2%) or failed to be processed (136 or 0.8%). Among the assigned bins, ‘Protein’ (1,581 or 9.0%) and ‘RNA’ (1002 or 5.7%) were the most abundant while bins ‘gluconeogenesis/glyoxylate cycle’ (3 or 0.02%) and ‘micro RNA, natural antisense etc’ (1 or 0.01%) the least (Figure 2.4).

2.4.5 Chromosomal locality or apple genome origin of the new transcripts

Chromosomal locality of the 8,144 transcripts identified in the first round of revision was inferred straightforwardly from their harboring MDCs in the apple genome Md-v1.0. It showed that 7,582 (93.1%) transcripts were located in MDCs anchored to one of the 17 chromosomes while 562 (6.9%) were found in unanchored MDCs (Figure 2.5). The apple genome origin of the 9,380 new transcripts that were identified from the unmapped reads in the second and third rounds of revisions (Figure 2.1) was evaluated by BLAST searches against Md-v1.0 and GenBank. The BLAST searches found that 9,375 of them had one or more significant hits ($E<10^{-10}$) in the apple genome Md-v1.0 (Table 2.5). Of the 9,375 of significant hits in the apple genome, 7,594 (81.0%) had the highest sequence identity greater than 98.0%, and the rest 1,781 (19.0%) ranged from 70.0% through 98.0%. Among the remaining five transcripts, two had significant hits of strawberry sequences in GenBank and three did not show any significant similarities with any sequences in GenBank (Table 2.5). These data suggested that 9,377 of the 9,380 new transcripts are of the apple genome origin. The three without any significant hits in GenBank are also likely of the apple genome origin (see Discussion).

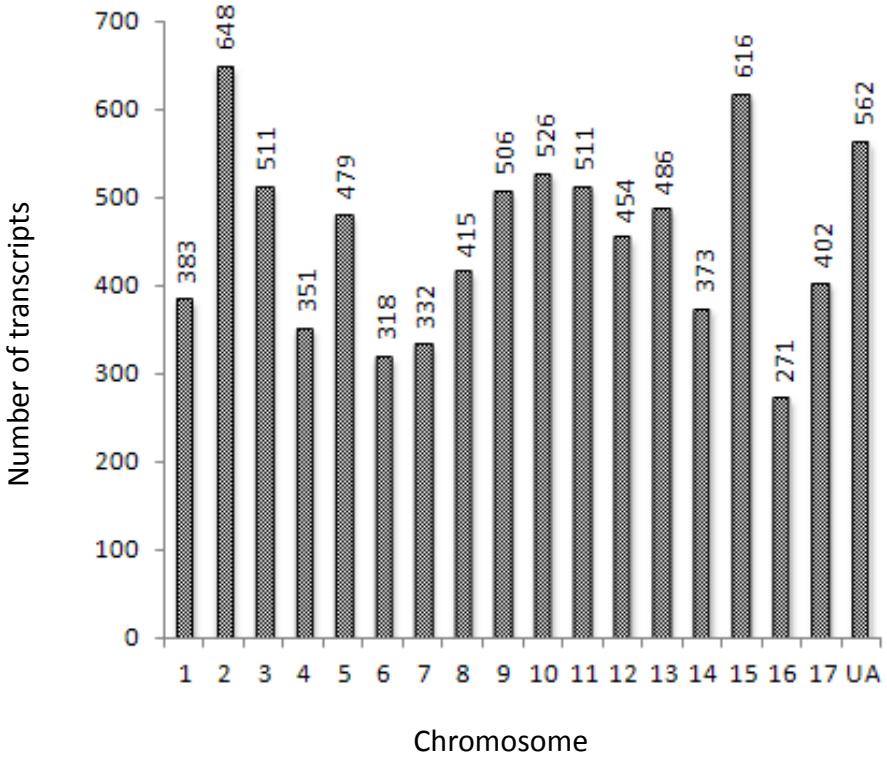


Figure 2.5 Chromosomal distribution of the 8,144 novel transcripts. *UA* unanchored in the apple genome

Table 2.5 The number of novel transcripts returned with one or more significant hits in BLAST searches [1]

Databases searched	Highest sequence identity (%)	E value	No. of transcripts	Percent
Md-v1.0	98.01-100.00	9.044E-18 to 0	7,594	80.96
Md-v1.0	90.01-98.00	6.732E-24 to 0	1,665	17.75
Md-v1.0	85.01-90.00	8.307E-22 to 0	75	0.80
Md-v1.0	70.01-85.00	8.893E-29 to 0	41	0.44
GenBank	74.53-78.10	1.66E-29 to 2.63E-71	2	0.02
Md-v1.0 & GenBank	NA	NA	3	0.03
Total			9380	100

[1] The cutoff is $E<10^{-10}$

2.5 Discussion

2.5.1 Identification of novel transcripts

The massive throughput of RNA-seq has been effective in identifying novel transcripts in annotated genomes (Trapnell et al. 2010; Zhang et al. 2010; Roberts et al. 2011). Through direct mapping of RNA-seq reads to the apple reference genome Md-v1.0, we identified 8,144 novel transcripts. These transcripts represent a fraction of the apple reference transcriptome missing in Md-v1.0-RT. Furthermore, by de novo assembling of reads that could not be mapped to Md-v1.0 in RNA-seq mapping, an additional 9,380 novel transcripts were uncovered. Taken together, these revisions represent another important section of the apple reference transcriptome. Collectively, we revealed 17,524 new transcripts in this study. Similar results were reported in other plant species. For example, mapping of RNA-seq reads had identified 5, 285 genes that were never annotated previously in the cucumber genome (Li et al. 2011). In barley, RNA-seq based gene annotations led to a report of 34,276 novel transcribed genomic regions (Olson et al. 2013).

The 9,380 new transcripts were identified from the reads that could not be mapped initially. To provide evidence that they were of apple origin rather than from other sources or contamination, the 9,380 new transcripts were BLAST searched against Md-v1.0 and GenBank databases. The searches found that 9,377 of them were returned with a significant hit(s) in apple (9,375) or strawberry (2) (Table 2.5), strongly suggesting that these transcripts are of the apple genome origin. However, the chromosomal locality of the 9,380 new transcripts could not be determined in this study. In order to locate them on chromosomes, a dedicated effort is necessary. It is highly likely that some of these new transcripts would come from genomic regions

currently missing in Md-v1.0. Determining their chromosomal locality would likely aid gap filling in the apple reference genome.

The three transcripts that had no significant hits in BLAST searches were also likely of the apple genome origin as they are not from contaminations of any known sources. Contamination by non-apple transcripts was indeed detected in the RNA-seq reads. In the BLAST searches, there were 56 transcripts assembled from the un-mapped reads that were identified nearly identical to genes of apple virus and other microbes, including apple chlorotic leaf spot virus, apple stem pitting virus, and apple green crinkle associated virus. These non-apple transcripts were removed in the process of finding this set of new transcripts (Steps 14-16, Figure 2.1). Since the 56 contaminant transcripts were found along with 9,377 apple transcripts, the likelihood for the three transcripts being of non-apple genome origin would be low. The chromosomal locality of these three transcripts will be determined together with the majority (9,377) of this group of novel transcripts.

A large number (9,627 or 55.2%) of the 17,524 new transcripts are unknown in MapMan gene ontology (code 35.2) (Figure 2.4). Several factors might have contributed to this observation. First, they were indeed new genes and there were no characterized orthologs in the databases. Second, they were un-translated regions (UTRs) disassociated with their coding regions of certain genes. Third, they were non-coding RNAs, which are common in plant genomes (Qi et al. 2013; Wu et al. 2013). Examining the 17,524 new transcripts showed that a CDS longer than 200bp could only be found in 8,629 of them. This strongly suggested that the majority of the remaining 8,895 represent either UTRs or non-coding RNAs, largely explaining the observation.

There were 74 genes derived from merging two or more MDPs (154 single MDPs in total) because of the bridging reads between them on the same strand. These

merges appeared to be similar to gene fusions that generate hybrid genes when chromosomal rearrangements bring two separate genes together. However, based on studies in plant genomes, such as Arabidopsis and rice, gene fusion is a rare and slow process (Nakamura et al. 2007). The observations of bridging reads between adjacent MDPs on the same stand were therefore unlikely caused by gene fusion, but by the imperfect gene prediction in Md-v1.0.

2.5.2 Improvement of the apple reference transcriptome

The apple reference genome Md-v1.0 has been available since 2010 (Velasco et al. 2010). Although the efforts are underway, Md-v2.0 has not been released thus far. We presented here an approach of repeated read mapping and transcript discovery using CLC Genomics Workbench to improve the original version of reference transcriptome Md-v1.0-RT. The major improvement is the identification of the 17,524 new transcripts. The improved reference transcriptome allowed 76.5±7.0% of the reads in the 14 samples mapped in RNA-seq mapping, representing a marked increase from 42.8±4.5%, the reads mapping rate associated with the original Md-v1.0-RT. Testing of the improved reference with two published datasets from other genotypes and/or tissue types also showed a notable improvement in coverage of reads mapping. In the first dataset (Krost et al 2012 and 2013), the coverage was increased from 46.6±7.1% to 82.3±2.7%. In the second dataset (Gusberti et al. 2013), it was improved from 37.4±7.7% to 62.5±9.3%. These results suggested that the improved apple reference transcriptome may be used in RNA-seq based studies involving tissues beyond fruit and genotypes beyond Golden Delicious.

The coverage of a reference transcriptome in mapping RNA-seq reads is affected by many factors, such as the purity of source mRNA, reads length, and

mapping parameters, especially the minimum length fraction and the minimum similarity. In this study, we set the minimum length fraction and the minimum similarity to 0.80 and 0.98, respectively. Testing of RNA-seq mapping using the six samples from Gusbeti et al (2013) (Table 2.2) against the original reference transcriptome Md-v1.0-RT showed that the 0.80/0.98 combination of parameters allowed mapping 28.8±5.9% of the reads uniquely (Table 2.4). This is equivalent to an 8.2-percentile reduction from the unique mapping rate (37.0±3.7%) for the same six samples reported in Gusberti et al (2013), where the CLC default mapping parameters (0.90 in the minimum length fraction and 0.80 in the minimum similarity) were used. The mapping parameters (0.80/0.98) are therefore probably more stringent than the CLC default settings in RNA-seq mapping, suggesting that the reference coverage was likely under-estimated in this study. Using Md-v1.0-RT, a read mapping rate of 65% was reported (Gapper et al. 2013). It was possible that the shorter read length (40 bases) and lower minimum similarity (0.95) used in the study might have contributed to the observation of a relatively high reads coverage for the original reference transcriptome Md-v1.0-RT.

In the three-round reads mapping and transcript discovery approach, we restricted RNA-seq reads to Golden Delicious, the source of the reference genome Md-v1.0 to avoid potential uncertainties from other genotypes. This might not be necessary given the high reads mapping rate (82.3±2.7%) from the two other genotypes (Krost et al. 2012 and 2013). Nevertheless, the approach appeared to be effective as it had led to an improved reference transcriptome that gave high coverage in mapping RNA-seq reads. The important difference of this approach from the de novo transcriptome assembly approach (Krost et al. 2012 and 2013; Zhang et al. 2012) is that the improved reference transcriptome was mostly built on the existing reference genome and transcriptome. Since much of the improvement (e.g. the 8,144 novel

transcripts) could be localized on chromosomes, the resulting transcriptome, if used, would readily allow studies to put findings under the context of genome.

The high reads mapping rates were obtained from samples under normal growth conditions. It remains to be seen whether or not the improved reference transcriptome would also provide a high coverage for reads mapping when samples were treated with biotic and abiotic stresses. In the six samples (Table 2.2) from Gusberti e al. (2013), samples (ERR313216, ERR313224, ERR313239) challenged by *V. inaequalis* appeared to have a lower read mapping rate (58.2±3.3%) than their non-challenged controls (ERR313217, ERR313225 and ERR313226) did (66.9±12.2%), but the difference was insignificant in t-test (P=0.2322). A noteworthy point is that there were 9,887 MDPs not included in the improved reference transcriptome due to zero reads mapped to their home MDCs. Although excluding these MDPs had no or little effect on mapping RNA-seq reads in this study, they might become relevant if their expression were highly specific to certain tissues, conditions, or growth and developmental stages. It might be necessary to include them in the reference transcriptome when the improved reference transcriptome is not satisfactory in mapping RNA-seq reads. Another important point is that the improved reference genome is just one step forward towards the complete apple reference transcriptome. Much more work remains to be continued, which includes, but not limited to, mRNA sequence backed precise annotation of all protein coding genes that include UTRs and alternative splicing variants, and non-coding RNAs.

2.6 Conclusion

We improved the current apple reference transcriptome Md-v1.0-RT using three rounds of read mapping and transcript discovery based on the RNA-seq data from fruit

of Golden Delicious at 14 stages of growth and development. The major improvement is the identification of 17,524 novel transcripts that either not annotated or missing in the current reference genome. The improved reference transcriptome considerably increased the RNA-seq mapping rates in the samples studied, including those from genotypes rather than Golden Delicious. The improvement represents a step forward towards a complete reference transcriptome in apple.

References

Chepelev I, Wei G, Tang QS, Zhao KJ (2009) Detection of single nucleotide variations in expressed exons of the human genome using RNA-Seq. Nucleic Acids Res 37:1-8 (e106)

Gapper NE, Rudell DR, Giovannoni JJ, Watkins CB (2013) Biomarker development for external CO_2 injury prediction in apples through exploration of both transcriptome and DNA methylation changes. AoB Plants 5: plt021; doi:10.1093/aobpla/plt021

Gasic K, Hernandez A, Korban SS (2004) RNA extraction from different apple tissues rich in polyphenols and polysaccharides for cDNA library construction. Plant Mol Biol Rep 22:437-438

Gusberti M, Gessler C, Broggini GAL (2013) RNA-Seq analysis reveals candidate genes for ontogenic resistance in *Malus-Venturia* pathosystem. PLoS ONE 8(11): e78457. doi:10.1371/journal.pone.0078457

Krost C, Petersen R, Lokan S, Brauksiepe B, Braun P, Schmidt E (2013) Evaluation of the hormonal state of columnar apple trees (*Malus x domestica*) based on high throughput gene expression studies. Plant Mol Biol 81:211-220

Krost C, Petersen R, Schmidt ER (2012) The transcriptomes of columnar and standard type apple trees (*Malus x domestica*) - a comparative study. Gene 498:223-230

Li P, Ponnala L, Gandotra N, Wang L, Si Y, Tausta SL, Kebrom TH, Provart N, Patel R, Myers CR, Reidel EJ, Turgeon R, Liu P, Sun Q, Nelson T, Brutnell TP (2010) The developmental dynamics of the maize leaf transcriptome. Nat Genet 42:1060-1067

Li Z, Zhang Z, Yan P, Huang S, Fei Z, Lin K (2011) RNA-Seq improves annotation of protein-coding genes in the cucumber genome. BMC Genomics 12:540

Lister R, O'Malley RC, Tonti-Filippini J, Gregory BD, Berry CC, Millar AH, Ecker JR (2008) Highly integrated single-base resolution maps of the epigenome in Arabidopsis. Cell 133:523-536

Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods 5:621-628

Nakamura Y, Itoh T, Martin W (2007) Rate and polarity of gene fusion and fission in *Oryza sativa* and *Arabidopsis thaliana*. Mol Biol Evol 24:110-121

Olson A, Klein RR, Dugas DV, Lu Z, Regulski M, Klein PE, Ware D (2013) Expanding and vetting sorghum bicolor gene annotations through transcriptome and methylome sequencing. the plant genome. doi: 10.3835/plantgenome2013.08.0025 (Posted online 13 Sept. 2013)

Ong WD, Voo L-YC, Kumar VS (2012) De novo assembly, characterization and functional annotation of pineapple fruit transcriptome through massively parallel sequencing. PLoS ONE 7:e46937

Qi X, Xie S, Liu Y, Yi F, Yu J (2013) Genome-wide annotation of genes and noncoding RNAs of foxtail millet in response to simulated drought stress by deep sequencing. Plant Mol Biol 83:459-473

Roberts A, Pimentel H, Trapnell C, Pachter L (2011) Identification of novel transcripts in annotated genomes using RNA-Seq. Bioinformatics 27:2325-2329

Ruttink T, Sterck L, Rohde A, Bendixen C, Rouzé P, Asp T, Van de Peer Y, Roldan-Ruiz I (2013) Orthology guided assembly in highly heterozygous crops: creating a reference transcriptome to uncover genetic diversity in *Lolium perenne*. Plant Biotechnol J 11: 605-617

Sun T, Germain A, Giloteaux L, Hammani K, Barkan A, Hanson MR, Bentolila S (2013) An RNA recognition motif-containing protein is required for plastid RNA editing in Arabidopsis and maize. Proc Natl Acad Sci 110:E1169-E1178

Suzuki H, Yu J, Ness S, O'Connell M, Zhang J (2013) RNA editing events in mitochondrial genes by ultra-deep sequencing methods: a comparison of cytoplasmic male sterile, fertile and restored genotypes in cotton. Mol Genet Genomics 288:445-457

Thimm O, Blasing O, Gibon Y, Nagel A, Meyer S, Kruger P, Selbig J, Muller LA, Rhee SY, Stitt M (2004) MAPMAN: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. Plant J 37:914-939

Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol 28:511-U174

Velasco R, Zharkikh A, Affourtit J, Dhingra A, Cestaro A, Kalyanaraman A, Fontana P, Bhatnagar SK, Troggio M, Pruss D, Salvi S, Pindo M, Baldi P, Castelletti S, Cavaiuolo M, Coppola G, Costa F, Cova V, Dal Ri A, Goremykin V, Komjanc M, Longhi S, Magnago P, Malacarne G, Malnoy M, Micheletti D, Moretto M, Perazzolli M, Si-Ammour A, Vezzulli S, Zini E, Eldredge G, Fitzgerald LM, Gutin N, Lanchbury J, Macalma T, Mitchell JT, Reid J, Wardell B, Kodira C, Chen Z, Desany B, Niazi F, Palmer M, Koepke T, Jiwan D, Schaeffer S, Krishnan V, Wu C, Chu VT, King ST, Vick J, Tao Q, Mraz A, Stormo A, Stormo K, Bogden R, Ederle D, Stella A, Vecchietti A, Kater MM, Masiero S, Lasserre P, Lespinasse Y, Allan AC, Bus V, Chagne D, Crowhurst RN, Gleave AP, Lavezzo E, Fawcett JA, Proost S, Rouze P, Sterck L, Toppo S, Lazzari B, Hellens RP, Durel C-E, Gutin A, Bumgarner RE, Gardiner SE, Skolnick M, Egholm M, Van de Peer Y, Salamini F, Viola R (2010) The genome of the domesticated apple (*Malus x domestica* Borkh.). Nat Genet 42:833-839

Wang A, Xu K (2012) Characterization of two orthologs of REVERSION-TO-ETHYLENE SENSITIVITY1 in Apple. J Mol Biol Res 2:24-41

Wilhelm BT, Landry JR (2009) RNA-Seq-quantitative measurement of expression through massively parallel RNA-sequencing. Methods 48:249-257

Wilhelm BT, Marguerat S, Goodhead I, Bahler J (2010) Defining transcribed regions using RNA-seq. Nat Protocols 5:255-266

Wu H-J, Wang Z-M, Wang M, Wang X-J (2013) Widespread long noncoding rnas as endogenous target mimics for microrna in plants. Plant Physiol 161:1875-1884

Zenoni S, Ferrarini A, Giacomelli E, Xumerle L, Fasoli M, Malerba G, Bellin D, Pezzotti M, Delledonne M (2010) Characterization of Transcriptional Complexity during Berry Development in *Vitis vinifera* Using RNA-Seq. Plant Physiol 152:1787-1795

Zhang G, Guo G, Hu X, Zhang Y, Li Q, Li R, Zhuang R, Lu Z, He Z, Fang X, Chen L, Tian W, Tao Y, Kristiansen K, Zhang X, Li S, Yang H, Wang J, Wang J (2010) Deep RNA sequencing at single base-pair resolution reveals high complexity of the rice transcriptome. Genome Res 20:646-654

Zhang Y, Zhu J, Dai H (2012) Characterization of transcriptional differences between columnar and standard apple trees using RNA-seq. Plant Mol Biol Rep 30:957-965

Zhong S, Joung J-G, Zheng Y, Chen Y-r, Liu B, Shao Y, Xiang JZ, Fei Z, Giovannoni JJ (2011) High-throughput illumina strand-specific rna sequencing library preparation. Cold Spring Harbor Protoc 2011: doi:10.1101/pdb.prot5652

CHAPTER 3

A CO-EXPRESSION GENE NETWORK REGULATING ACIDITY IN DEVELOPING APPLE FRUIT

3.1 Abstract

Apple fruit acidity, which affects fruit overall taste and flavor to a large extent, is primarily determined by the concentration of malic acid. Previous studies demonstrated that the major QTL *Ma (malic acid*) on chromosome 16 is largely responsible for fruit acidity variations in apple. Recent advances suggested that a natural mutation led premature stop codon in one of the two aluminum-activated malate transporter (ALMT)-like genes (called *Ma1*) is the gene underlying *Ma*. However, these findings do not explain the developmental changes of fruit malate levels in a given genotype. Using RNA-seq data from the fruit of 'Golden Delicious' taken at 14 developmental stages, we characterized their transcriptomes in groups of high (12.2±1.6 mg/g fw), mid (7.4±0.5 mg/g fw) and low (5.4±0.4 mg/g fw) malate concentrations. Detailed analyses showed that a set of 3,066 genes (including *Ma1*) were expressed not only differentially (P_{FDR}<0.05) between the high and low malate groups, but also in significant (P<0.05) correlation with malate concentrations. The 3,066 genes fell in 648 MapMan (sub-) bins or functional classes, and 19 of them, which encompassed 363 genes, were significantly (P_{FDR}<0.05) co-enriched or co-suppressed in a malate dependent manner. Network inferring (r>0.94) using the 363 genes led to a major co-expression gene network of 286 genes. Based on the putative functions of the 19 functional classes and the core members in the gene network, we suggest that: (1) the pathways of the malate-pyruvate interconversion, photosynthesis,

mitochondrial electron transport and amino acid degradation are involved in the variation of malate levels in developing fruit; and (2) the co-expression gene network not only involves symplastic signal transduction, transcription regulation and post-translational modification, but also the role of apoplast.

3.2 Introduction

Acidity, which affects fruit overall taste and flavor, has long been considered one of the most important quality attributes in fruits, such as apple, grape and tomato. As such, titratable acidity and pH are routinely measured to evaluate fruit acidity levels for advancing selections in fruit breeding. Many organic acids contribute to acidity levels in fruit, but malic acid (or its conjugate base malate) and citric acid (or citrate) are considered the major contributors. Given their roles in determining fruit acidity, extensive efforts have been made to understand how malate and citrate accumulate in fruit cells, which has been reviewed recently in detail (Etienne et al. 2013; Sweetman et al. 2009).

In apple, acidity in mature fruit is primarily determined by the concentration of malic acid as it usually makes up approximately 98% of total acidity (Hulme and Wooltorton 1957). Obviously, biochemical and physiological processes that affect malate accumulations are relevant in determining fruit acidity levels. These processes may include malate synthesis, degradation, intracellular transport, and storage (Beruter 2004; Etienne et al. 2013; Sweetman et al. 2009). Malate is primarily synthesized in cytosol through glycolysis of hexoses derived from sorbitol and sucrose, which are imported from leaves. It could also be synthesized through pathways pertaining to photosynthesis in chloroplast, the tricarboxylic acid (TCA) cycle in mitochondrion, and the glyoxylate cycle in glyoxysome at varying fruit growth and developmental

stages. For malate degradation, the main pathways are the gluconeogenesis in cytosol and the TCA cycle in mitochondrion. In addition to synthesis and degradation, intracellular transport of malate into and out of the vacuole is an integral part of the network for malate homeostasis in cells as vacuole is the major repository of malate.

To understand these biological processes important for fruit acidity in apple, genes and/or enzymes involved in malate synthesis and degradation are logical targets of investigations. However, studies on the relevance of these genes or enzymes in determining acidity levels had drawn inconsistent conclusions. On one hand, no difference in the activity of phosphoenolpyruvate carboxylase (PEPC), NAD-dependent malate dehydrogenase (NAD-MDH) or NADP-dependent malic enzyme (NADP-ME), the key enzymes in malate metabolism, was found between a low acid variety 'Usterapfel' and its high acid mutant (Beruter 2004), suggesting a limited role, if any, of these enzymes in determining apple fruit acidity. On the other hand, profiling the expression patterns and their corresponding enzyme activities for genes MdPEPC (EU315246, for PEPC) and MdcyME (DQ280492, for NADP-ME) underscored that there were differences between low and high acid genotypes (Yao et al. 2009), implicating their roles in fruit acidity. In addition, the functionality of NAD-MDH (DQ221207) had been demonstrated in malate synthesis in apple (Yao et al. 2011).

Although the study (Beruter 2004) did not find significant differences between the low acid variety 'Usterapfel' and its high acid mutant in the catalytic activity for enzymes PEPC, NAD-MDH and NADP-ME, it demonstrated that the uptake of [^{14}C] malate was significantly lower in excised tissue of 'Usterapfel' than in that of the mutant. This suggested that the low malate content in 'Usterapfel' fruit was the result of a restricted ability to transport malate into its vacuoles and then maintain the malate levels in them. Indeed, several vacuolar transporters, such as the vacuolar pumps, e.g.

V-ATPase (Schumacher and Krebs 2010) and MdVHA-A (EF128033, for subunit A of vacuolar H+-ATPase)(Yao et al. 2009), tonoplast dicarboxylate transporter, e.g. AttDT (Emmerlich et al. 2003), and members of the ALMT1 (aluminum-activated malate transporter1) family proteins (Barbier-Brygoo et al. 2011), e.g. AtALMT9 (Kovermann et al. 2007) and AtALMT6 (Meyer et al. 2011), had been implicated of critical roles in maintaining the homeostasis of malate in plant cells. The most important and direct evidence of the role of vacuolar transporters in determining malate levels in apple came from the isolation of *Ma* (*malic acid*) (Bai et al. 2012; Khan et al. 2013), a major gene or QTL controlling fruit acidity on linkage group 16 (Kenis et al. 2008; Liebhard et al. 2003; Maliepaard et al. 1998; Xu et al. 2012). The *Ma* locus encodes two ALMT1 like genes, called *Ma1* and *Ma2*, and a single nucleotide mutation that led to a premature stop codon in *Ma1* was attributed to the low acidity phenotype in apple(Bai et al. 2012).

In developing apple fruit of a given variety, the malate concentrations could vary by three folds between the highest and the lowest. The change commonly begins with a rapid increase in the first 4-5 weeks after full-bloom and then progressively decreases through maturity (Beruter 2004; Hulme and Wooltorton 1957; Ulrich 1970; Zhang et al. 2010). Although the *Ma* locus largely governs fruit acidity variations in diverse apple varieties, it does not explain the developmental changes of fruit acidity in a given genotype as the genotype at the *Ma* locus never changes. We hypothesized that developmental variations in apple malate levels were caused by a set of genes that were expressed differentially and in a malate dependent manner. The objectives of this study were 1) to identify the genes that were differentially expressed in developing fruit of significant variations at malate levels, 2) to identify functional classes or MapMan (sub-) bins that were either co-enriched or co-suppressed in a malate dependent manner and construct a co-expression gene network of these genes.

3.3 Materials and Methods

3.3.1 Plant materials and malate quantification

The fruit samples of 'Golden Delicious' (GD) that were collected in a previous study (Wang and Xu 2012) were used in this work. Briefly, the fruit samples were taken at 14 time points from 1 week after full-bloom (WAF01) through harvest (WAF20) in 2010 and were immediately frozen in liquid nitrogen and stored at -80℃. For WAF01-03 (3 time points), the whole fruit were pooled with the pedicel removed, and there were at least ten fruit for each sampling time point. For the rest 11 time points WAF04-20 only the cortex tissues were used and at least five fruit were pooled for each time point. In addition, three biological replicates at WAF20 were also prepared from 1, 2 and 2 fruit that were ground independently. As a result, a total of 17 fruit samples were prepared for the 14 time points (1 pooled sample for each time point of WAF01-20 plus 3 additional replicates for WAF20).

Malate concentrations of the 14 pooled fruit samples were determined on an Agilent 7890A GC/5795C MS (Agilent Technology, Palo Alto, CA, USA) with three technical replicates from the pooled fruit samples. Organic acids were extracted and derivatized following a protocol described previously (Lisec et al. 2006) with minor modifications. Briefly, homogenized apple fruit tissues of 100 mg (per sample per replication) were extracted in 1.4 ml of 75% methanol with 600 ppm ribitol added as internal standard. After fractionation of non-polar metabolites into chloroform, 2.5 µl of the polar phase were transferred into a 2.0 ml Eppendorf tube and were dried under vacuum without heating and then derivatized with 40 ul each of methoxyamine hydrochloride an N-methyl-N-trimethylsilyl-trifluoroacetamide (MSTFA) sequentially. To measure metabolites, 1 ul of the derivatized sample was injected at 230°C in

splitless mode with helium carrier gas flow set to 1 ml/min. Chromatography was performed on a DB-5MS capillary column (20 m × 0.18 mm × 0.18 μm) with a 5 m Duraguard column (Agilent Technology). The temperature program was isothermal at 70°C for 2.471 min, followed by a 10.119°C/min ramp to 330°C and a final 2.471 min heating at 330°C. Cooling was performed as fast as possible. The system was then temperature equilibrated at 70°C for 5 min before the next injection. Mass spectra were collected at 5.6 scans s−1 over an m/z 50–600 scanning range. The transfer line temperature and the ion source temperature were set to 250 and 230°C, respectively. Metabolites were identified by comparing fragmentation patterns with those in a mass spectral library generated on our GC/MS system and an annotated quadrupole GC–MS spectral library downloaded from the Golm Metabolome Database (http://csbdb.mpimp-golm.mpg.de/csbdb/gmd/msri/gmd_msri.html) and quantified based on standard curves generated for each metabolite and internal standard.

3.3.2 RNA isolation and strand specific RNA-seq library construction and sequencing

For total RNA isolation, ground fruit tissue of 2g (young fruit)-3g (mature or near mature fruit) were used following a previous protocol (Gasic et al. 2004) with modifications, i.e. 1ml Sarkosyl of 20% (w/v) was added to the extraction buffer before tissue homogenization. The isolated total RNA was stored in EB buffer (Qiagen, Germantown, MD) with addition of 1× Ambion RNAsecure (Invitrogen/Life Technologies, Carlsbad, CA). Activation of RNAsecure was carried out by incubating the samples at 60°C (in a water bath) for 10 min and then immediately put on ice. Evaluation of RNA quantity and quality was performed using Nanodrop 1000 (Thermo Scientific, Waltham, MA) and Bioanalyzer 2100 with RNA 6000 Nano Chip (Agilent, Santa Clara, CA) and electrophoresis of a 2% agarose gel (using 1/10 RNA

dilutions in EB buffer with 1× Ambion RNA secure). Immediately prior to mRNA isolation, the RNA samples were treated with DNase I (amplification grade, Invitrogen) at 37°C for 30 min followed by heat inactivation at 65°C for 15 min. The procedure for RNA-seq library construction was described previously (Bai et al. 2014). Briefly, we used NEBNext Poly(A) mRNA Magnetic Isolation Module and NEBNext Ultra Directional RNA Library Prep Kit for Illumina (New England Biolabs, Ipswich, MA) for mRNA isolation and stand specific RNA-seq library preparation. The starting amount of total RNA for each library was 5μg and the manufacturer's protocols were followed with minor modifications (Bai et al. 2014). The 14 libraries constructed from pooled fruit samples were multiplexed with 60 ng each for single-end sequencing of 101 bases without replication in one lane of Illumina HiSeq 2000 (Illumina, San Diego, CA) at the Cornell University Biotechnology Resource Center (Ithaca, NY). The three libraries constructed from the three WAF20 replicates were multiplexed with other samples and sequenced similarly in another lane.

3.3.3 RNA-seq data analysis

Seventeen sequence files with a total of 214.2 million (data not shown) raw reads were generated by the Illumina pipeline in software CASAVA v1.8 in Sanger FASTQ format (available under NCBI SRA experiment number SRX392051), but only those (199.8 million, 93.3±1.2% of the total raw reads) that passed the chastity filter (i.e. no more than one base call in the first 25 cycles has a chastity higher than 0.6) in the pipeline were used for further analysis (Table 3.1).

Table 3.1 Overview of RNA-seq reads mapping

Sample Name	Reads passed CASAVA1.8	Clean and high quality reads	Total reads mapped	Rate of total reads mapped (%)	Unique reads mapped	Rate of unique reads mapped (%)	Overall rate of mapped reads (%)
WAF01	13,212,078	11,345,227	9,116,208	80.4	7,633,852	67.3	69.00
WAF02	11,577,420	10,705,459	8,630,679	80.6	7,207,324	67.3	74.55
WAF03	13,064,054	12,964,570	10,517,983	81.1	8,792,486	67.8	80.51
WAF04	12,902,501	12,837,350	10,373,060	80.8	8,675,165	67.6	80.40
WAF05	13,594,031	12,970,807	10,493,887	80.9	8,781,681	67.7	77.19
WAF06	13,063,572	11,704,599	9,402,917	80.3	7,819,313	66.8	71.98
WAF08	11,692,039	8,214,001	6,422,760	78.2	5,310,767	64.7	54.93
WAF10	12,036,125	11,806,215	9,512,624	80.6	7,931,225	67.2	79.03
WAF12	10,788,214	10,602,253	8,515,180	80.3	7,142,042	67.4	78.93
WAF14	11,343,438	10,497,848	8,168,955	77.8	6,831,534	65.1	72.01
WAF16	10,698,751	9,384,232	7,121,602	75.9	5,934,340	63.2	66.56
WAF18	11,701,713	9,661,120	7,215,878	74.7	6,029,985	62.4	61.67
WAF19	10,729,005	9,813,882	7,750,642	79.0	6,483,387	66.1	72.24
WAF20*	*10,858,938*	*9,356,135*	*7,323,393*	*78.19*	*6,091,529*	*65.01*	*66.85*
WAF20-P	12,124,220	11,885,016	9,336,067	78.6	7,787,968	65.5	77.00
WAF20-I	10,181,395	7,585,469	5,872,330	77.4	4,835,327	63.7	57.68
WAF20-II	8,731,735	7,221,069	5,642,027	78.1	4,714,082	65.3	64.62
WAF20-III	12,398,400	10,732,987	8,443,147	78.7	7,028,740	65.5	68.10
Total	199,838,691	179,932,104	142,535,946		118,939,218		
Mean	11,755,217	10,584,241	8,384,467	79.02	6,996,425	65.91	70.96
SD	1,269,947	1,773,230	1,538,605	1.87	1,297,956	1.68	7.88

*Mean of the four samples from WAF20 (not included in calculation of column totals and means)

Data analyses were performed using CLC Genomics Workbench (CLC GW) v6.5 (CLCBio, Cambridge, Massachusetts). To remove reads derived from rRNA, the ribosomal RNA database SILVA Release 115 (http://www.arb-silva.de/) was downloaded and used as reference sequences. Mapping of the reads to the rRNA references was conducted using the minimum length fraction of 1.0 and the minimum similarity of 0.95. The reads that could not be mapped to the reference SILVA Release 115 were collected and then filtered by quality (using the CLC GW default settings) to further remove low quality reads and/or bases (Table 3.1). For mapping of RNA-seq reads against the improved apple reference transcriptome (Bai et al. 2014), the minimum length fraction of 0.8 and the minimum similarity of 0.98 were used (our empirical sequence identity threshold often effective in differentiating paralog sequences in the apple genome). The improved reference transcriptome is available at the Genome Database for Rosaceae (GDR), which includes 53,654 of the 63,541 genes or MDPs (MDPs are the three letters prefixing in apple gene IDs) predicted originally and 17,524 novel transcripts (Bai et al. 2014). For convenience, hereafter the 17,524 novel transcripts will be referred to 'gene' and 'MDP0000' (e.g. MDP0000252114) will be abbreviated to 'M' (e.g. M252114).

For expression analysis, the count of reads mapped for a given gene was expressed as RPKM (reads per kilobase exon model per million mapped reads) (Mortazavi et al. 2008). Genes were considered expressed if their RPKM values were greater than 0.3 as similarly defined previously (ES et al. 2001). For time point WAF20, the mean RPKM value of the four samples was used.

The cutoff for a gene said to be expressed differentially between malate groups is $P_{FDR}<0.05$ in the Baggerly's test (Ferguson 1984). In the subsequent gene co-enrichment or co-suppression analysis, K-mean clustering and network construction, RPKM values were square root transformed.

3.3.4 Identification of genes and MapMan bins associated with the variation of malate concentrations

We define that malate associated genes are those that meet the following two criteria: 1) expressed significantly (P_{FDR}<0.05) differently between high and malate samples in the Baggerly's test (Ferguson 1984) (conducted using CLC GW), and 2) expressed in significant correlation (P_r<0.05) with malate concentrations and/or the expression of *Ma1* (Bai et al. 2012) in the 14 sampling time points (conducted using MS Excel). Association of MapMan (sub-)bins or functional classes with malate was established using the PageMan tool (Usadel et al. 2006) embedded in the MapMan software, where the significant threshold P_{FDR}<0.05 was set in Wilcoxon rank sum test (Wilcoxon 1945) that was followed by the Benjamini-Hochberg false discovery rate (FDR) correction (Benjamini and Hochberg 1995).

To assign the MapMan (sub-) bins or functional classes, the 71,178 genes in the improved apple reference transcriptome (Bai et al. 2014) was conducted by BLAST search of multiple databases using the web-based search tool Mercator (http://mapman.gabipd.org/web/guest/mercator) (Lohse et al. 2014). The databases searched include: TAIR-Arabidopsis TAIR proteins (release 10), PPAP-SwissProt/UniProt Plant Proteins, CHLAMY-JGI Chlamy release 4 Augustus models, ORYZA-TIGR5 rice proteins, KOG: Clusters of orthologous eucaryotic genes database (KOG), CDD- conserved domain database, and IPR-interpro scan. The output file of Mercator lists the best hits in databases and assigns MapMan's (sub-) bins (Thimm et al. 2004) for each input sequences if possible, which were used as the 'mapping file' required by the PageMan tool. The 'experiment file' contained only the malate associated genes (met with the two criteria above) and their RPKM values with square root transformation.

3.3.5 Gene clustering and co-expression network inferring

Genes in the MapMan (sub-) bins that were associated with malate variations were clustered by K-means using MeV4.9 (Saeed et al. 2006) with default parameters except for the number of clusters, which was set to six. Network inferring from these genes was conducted using the Cytoscape Network Inference (Cyni) tool based on the Pearson correlation r>0.94, an optimal threshold in this study. For network analysis, the Cytoscape plugin NetworkAnalyzer (Assenov et al. 2008) was used.

3.3.6 qRT-PCR analysis

The same set of total RNA samples used for RNA-seq libraries were used for qRT-PCR. Three independent reverse transcription reactions were carried out for each sample with 2 μg of total RNA using the Superscript III RT (Invitrogen, Carlsbad, CA). The resulting first strand cDNA was diluted by 5 folds, and then used as templates for qRT-PCR, in which an apple actin gene (EB136338) served as a reference using primers 5'GGCTGGATTTGCTGGTGATG and 5'TGCTCACTATGCCGTGCTCA. The three targeted genes were M252114 (*Ma1*), M196894 and M127123, and their primer sequences (in pair) were 5'CGTCATGGTGTCTGGAACAT and 5'CTCCATGGCAAAAACCTGTC, 5'GCATCACGAAGAAGACGATG and 5'TTCTTGCCGTGAATCAACAA, and 5'GTGTGTGGGAATGAGGTGGA and 5'GTTTGATGGTGCTGGGTGAA, respectively. qRT-PCR was performed with the LightCycler 480 Real-Time PCR System (Roche, Indianapolis, IN) as described previously (Bai et al. 2012). Expression quantification and data analysis were performed via LightCycler 480 Software (Version 1.5) using the comparative cycle threshold method (Pfaffl 2001).

3.4 Results

3.4.1 Determination of malate concentrations in fruit and grouping

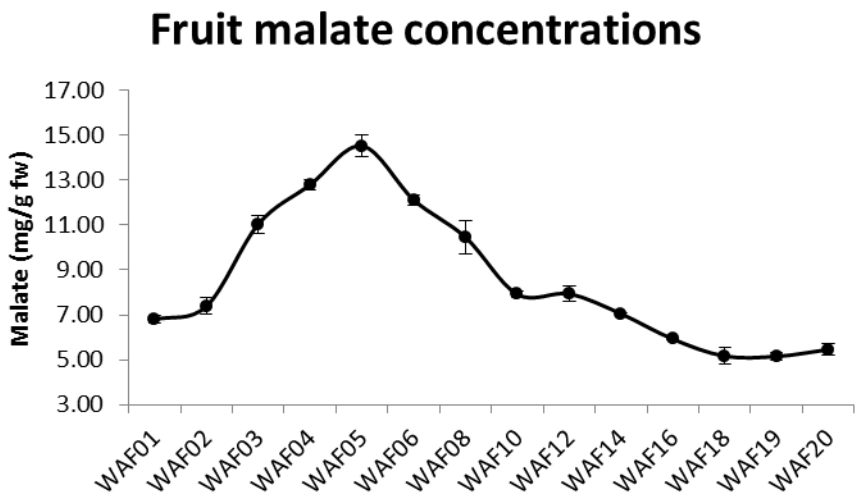
Fruit malate concentrations were measured by GC-MS at 14 stages from one week after full bloom (WAF01) to harvest (WAF20). The malate concentrations increased slowly from WAF01 (6.79±0.16mg/g fw) to WAF02 (7.39±0.38mg/g fw), but rapidly from WAF02 through WAF05 (14.52±0.48 mg/g fw) when the peak was reached (Figure 3.1A). Decline in malate concentration began at WAF06 and continued until harvest (WAF20) although there were minor upticks at WAF12 and WAF20. The lowest malate concentration was recorded at WAF19 (5.15 mg/g fw), equivalent to 35.5% of the peak at WAF05.

Based on the overall malate concentrations, the samples from the 14 time points were categorized into three groups: 1) high malate of five samples WAF03-WAF6 and WAF08, 2) mid malate of another five WAF01, WAF02, WAF10, WAF12 and WAF14, and low malate of four WAF16 and WAF18-WAF20 (Figure 3.1B). ANOVA analysis indicated that the differences in group means were significant between high and mid malate groups (P=1.8E-12, z test) as well as between the mid and low malate groups (P=7.8E-12, z test).

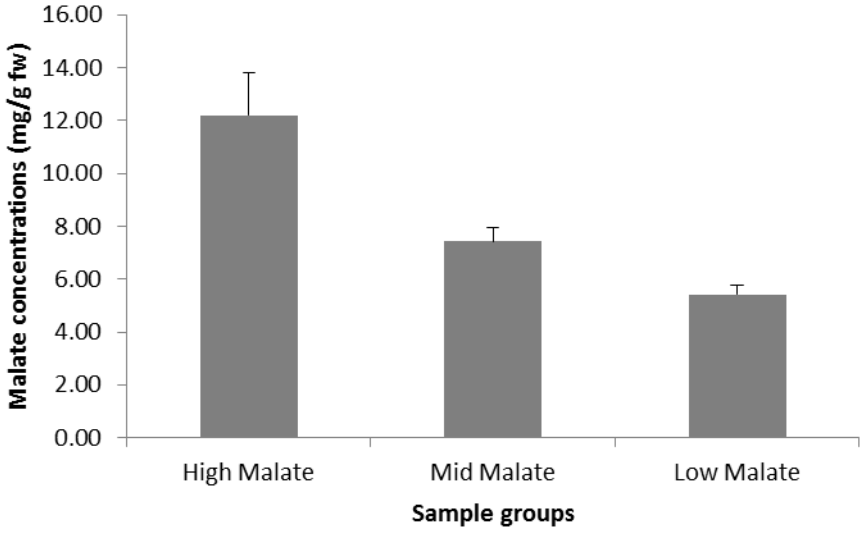
3.4.2 Expression analysis in groups of varying malate concentrations

Figure 3.1 Malate concentrations in developing fruit of Golden Delicious (GD). Standard deviations were shown with the error bars. **A.** GC-MS measurements of malate concentrations (mg/g fw) in fruit from one week after full-bloom (WAF01) through harvest (WAF20). **B.** Mean malate concentrations of fruit groups of high (WAF03-06 and WAF08), mid (WAF01-02, WAF12-16) and low (WAF16-20) malate.

A


Fruit malate concentrations
Malate (mg/g fw)
17.00
15.00
13.00
11.00
9.00
7.00
5.00
3.00
WAF01
WAF02
WAF03
WAF04
WAF05
WAF06
WAF08
WAF10
WAF12
WAF14
WAF16
WAF18
WAF19
WAF20

B


Malate concentrations (mg/g fw)
16.00
14.00
12.00
10.00
8.00
6.00
4.00
2.00
0.00
High Malate
Mid Malate
Low Malate
Sample groups

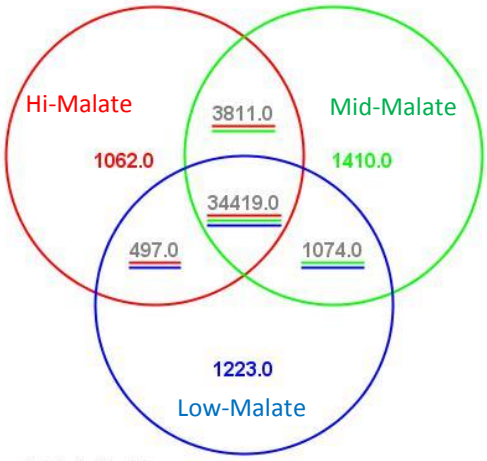
Gene expression analyses were conducted based on an improved apple reference transcriptome, which includes 53,654 genes (MDPs) and 17,524 new transcripts that were identified from RNA-seq reads unmapped to the original apple reference transcriptome (Bai et al. 2014). After removing reads derived from rRNA and those of low quality, the total reads input for mapping was 179.9 million, and the total mapped and uniquely mapped reads were 142.5 million (79.0%) and 118.9 million (65.9%), respectively (Table 3.1). The mean mapped reads per sample were 8.4±1.5 million in total and 7.0±1.3 million in unique.

Overall, there were 50,700 genes expressed (i.e. with an RPKM greater than 0.3 at least in one of the 14 time points, data not shown). However, if the mean RPKM values in the three groups were used, the number of expressed genes (i.e. with an RPKM greater than 0.3 at least in one of the three groups of high, mid and low malate) was 43,496 (Figure 3.2A). In individual malate groups, there were 39,789, 40,714 and 37,213 genes expressed in the high, mid and low malate groups, respectively. The genes that were uniquely expressed were 1,062, 1,410 and 1223 in the high, mid and low malate groups, respectively (Figure 3.2A). In terms of reads for the expressed genes, there were 8.6±1.6 million mapped total gene reads per time point in high malate group, 8.0±0.5 million in mid malate group, and 6.7±0.3 million in the low malate group (Figure 3.2B).

3.4.3 Identification of genes associated with variations of malate concentrations

To identify genes putatively responsible for the variations in malate concentrations in developing fruit, we conducted two comparisons of expression patterns: the first was between the high and low malate groups, which identified 4,041 genes or transcripts expressed significantly differentially ($P_{FDR}<0.05$) based on the Baggerly's te

A



B

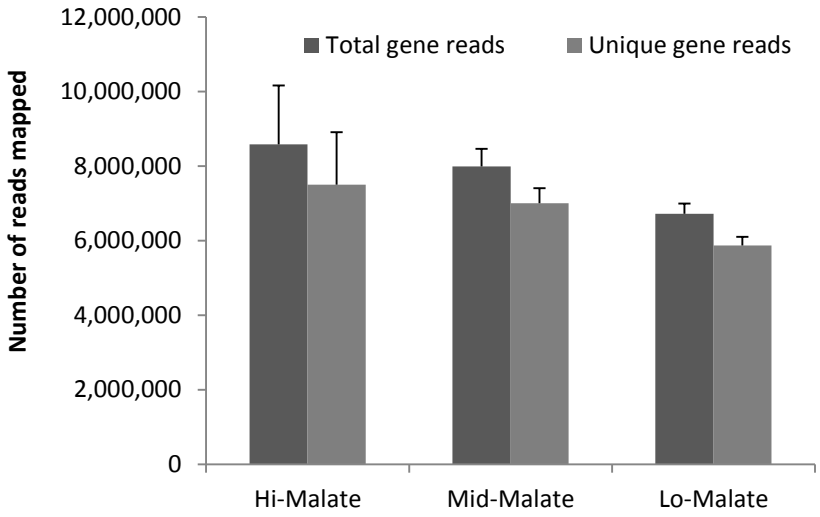


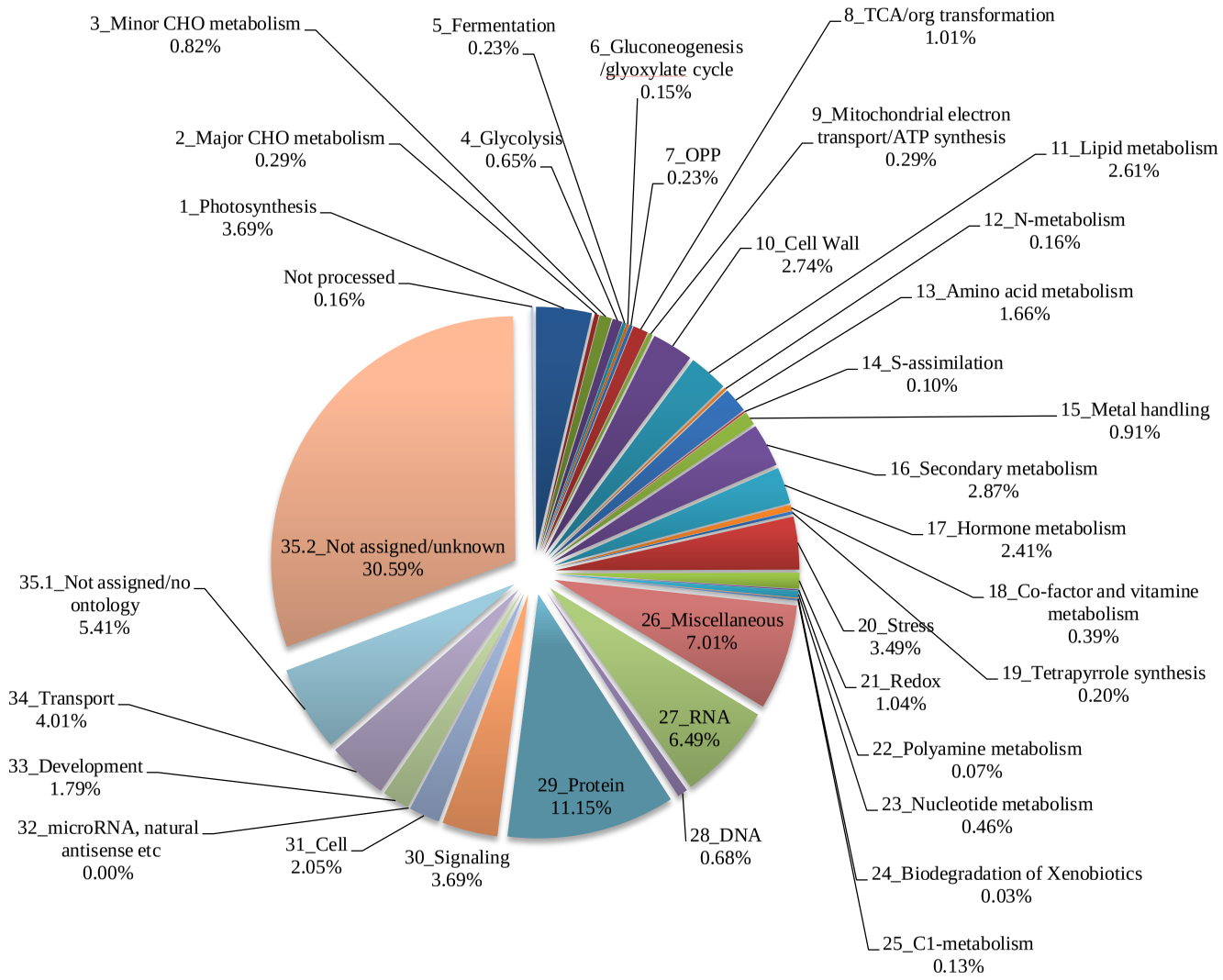
Figure 3.2 Overview of RNA-seq data analysis in the fruit groups of high, mid and low malate. **A.** Venn diagram representation of the number of genes expressed (RPKM>0.3). **B.** Mean number of the mapped reads (per sample). Standard deviations were shown with the error bars.

(Ferguson 1984). The second was among the 14 individual time points, which was started with a large set of 22,870 genes that showed minimal difference between any of the two time points, i.e. RPKM absolute difference>3 and RPKM absolute fold change >1.2. The genes in the large set were each calculated for their correlations with malate concentrations as well as the expression of *Ma1*. This identified 7,150 genes to be significantly ($P_{r>0.538}$<0.05) correlated with either malate concentrations and/or the expression of *Ma1*. The correlation coefficient (r) between malate and *Ma1* was r=0.572 ($P_{r=0.572}$<0.05), indicating that the expression of *Ma1* was significantly correlated with malate variations in developing fruits. Combining the two lists of genes 4,041 and 7,150 found 3,066 genes in common (including *Ma1*). These genes were not only expressed differentially (P_{FDR}<0.05) between high and low malate groups, but also significantly ($P_{r>0.538}$<0.05) correlated with malate concentrations (2180 or 71.1%), or with the expression of *Ma1* (100 or 3.3%), or both (786 or 25.6%).

MapMan gene ontology analysis showed that 1,949 (63.6%) of 3066 genes could be assigned to one or more MapMan Bins while 1,117 (36.4%) could not (Figure 3.3). Bins '29_protein' of 342 (11.5%) genes, '26_miscellaneous' of 215 (7.0%) and '27_RNA' of 199 (6.5%) were the largest bins among the assigned whereas '32_microRNA' of 0 (0%), '24_Biodegradation of Xenobiotics' of 1 (0.03%) and '22_Polyamine metabolism' of 2 (0.07%) were the smallest bins. In the unassigned, 174 (5.7%) were found of significant hits, but there were no ontology for them in MapMan; and the rest 938 (30.6%) were either unknown or failed to be processed in Mecator based database searches. Overall, the 3,066 genes were assigned into 648 MapMan (sub-) bins or functional classes (data not shown).

3.4.4 Identification of MapMan (sub-) bins expressed in a malate dependable manner

Figure 3.3 Distribution in MapMan bins of the 3,066 genes which were expressed not only significantly ($P<0.01$) in correlation with malate concentrations and/or the expression of *Ma1*, but also significantly ($P_{FDR}<0.05$) in difference between the high and low malate groups. The MapMan bins were coded with numbers 1-35, prefixing the name along with a dash.


3_Minor CHO metabolism
0.82%
5_Fermentation
0.23%
6_Gluconeogenesis
/glyoxylate cycle
0.15%
8_TCA/org transformation
1.01%
2_Major CHO metabolism
0.29%
4_Glycolysis
0.65%
7_OPP
0.23%
9_Mitochondrial electron
transport/ATP synthesis
0.29%
11_Lipid metabolism
2.61%
1_Photosynthesis
3.69%
12_N-metabolism
0.16%
10_Cell Wall
2.74%
Not processed
0.16%
13_Amino acid metabolism
1.66%
14_S-assimilation
0.10%
15_Metal handling
0.91%
16_Secondary metabolism
2.87%
17_Hormone metabolism
2.41%
35.2_Not assigned/unknown
30.59%
18_Co-factor and vitamine
metabolism
0.39%
35.1_Not assigned/no
ontology
5.41%
26_Miscellaneous
7.01%
20_Stress
3.49%
19_Tetrapyrrole synthesis
0.20%
21_Redox
1.04%
34_Transport
4.01%
27_RNA
6.49%
22_Polyamine metabolism
0.07%
33_Development
1.79%
29_Protein
11.15%
23_Nucleotide metabolism
0.46%
32_microRNA, natural
antisense etc
0.00%
28_DNA
0.68%
31_Cell
2.05%
30_Signaling
3.69%
24_Biodegradation of Xenobiotics
0.03%
25_C1-metabolism
0.13%

To determine if the expression of any functional classes or MapMan (sub-) bins were co-enriched or co-suppressed in three malate groups, the 3,066 genes were subjected to a MapMan Bin-wise Wilcoxon test. The resulting p values were adjusted according to the Benjamini and Hochberg (1995) false discovery rate control (FDR) and presented using the PageMan tool in MapMan, where the FDR p-values were further compressed by converting them into z-scores (a p-value of 0.05 is equal to a z-score of 1.96) (Figure 3.4). It showed that a total of 43 functional classes were expressed significantly higher or lower than the rest in a given malate group. Nineteen of the 43 functional classes fell into the two primary MapMan bins, i.e. '1_PS (photosynthesis)' of 8 and '29_protein' of 11. The rest (24) functional classes were in primary MapMan 12 bins with three each in '8_ TCA / org transformation', '20_stress', '27_RNA' and '30_signaling', two each in '13_amino acid metabolism', '16_secondary metabolism', '26_miscellaneous' and '35_not assigned', and one each in '9_mitochondrial electron transport / ATP synthesis', '10_cell wall', '21_redox' and '31_cell' (Figure 3.4).

There were two groups of functional classes to be excluded from further analysis. The first was the ten MapMan (sub-) bins in codes 1, 1.1, 1.1.1, 1.1.1.1, 1.3, 21, 29.2, 29.2.1, 35 and 35.2 (Figure 3.4) as they were either co-suppressed (in bins 35 and 35.2) or co-enriched (in the other eight) without discrimination in the three malate groups, indicating they were not causal for the variations of malate concentrations in developing fruit. The second was the 14 functional classes with MapMan bincodes 1.1.2, 8, 8.2, 13, 20.2, 26, 27, 27.3, 29, 29.2.1.1, 29.4, 29.4.1, 29.5 and 30.2 (Figure 3.4). These functional classes were in higher orders in the hierarchical nomenclature system of MapMan bins and their detections were likely caused by the presence of those in the (immediate) lower orders. For example, the detection of 1.1.2 was probably due to the presence of 1.1.2.2, and the detection of 8.2 due to the presence of 8.2.10 (Figure 3.4). Removing the second group is necessary to eliminate the

Figure 3.4 The functional classes (MapMan bins or sub-bins) that were significantly (P_{FDR}<0.05) co-enriched or co-suppressed in the fruit groups of high, mid and low malate. The FDR p-values were converted into z scores (a p-value of 0.05 equals a z-score of 1.96) and represented with a color scale from blue (co-suppressed) to white (expressed at normal levels) to red (co-enriched). The functional classes were annotated on the right panel and the numbers at the end of each annotation show the z scores in the low malate group.

High Malate
Mid Malate
Low Malate
1_photosynthesis (PS) 5.87
1.1_PS.lightreaction 4.66
1.1.1_PS.lightreaction.photosystem II 4.11
1.1.1.1_PS.lightreaction.photosystem II.LHC-II 3.21
1.1.1.2_PS.lightreaction.photosystem II.PSII polypeptide subunits 0.0
1.1.2_PS.lightreaction.photosystem I 0.0
1.1.2.2_PS.lightreaction.photosystem I.PSI polypeptide subunits 0.0
1.3_PS.calvin cycle 3.00
8_TCA/org transformation 3.43
8.2_TCA/org transformation.other organic acid transformations 3.28
8.2.10_TCA /org transformation.other organic acid transformations.malic 3.03
9_mitochondrial electron transport / ATP synthesis 2.40
10.6_cell wall.degradation -2.80
13_amino acid metabolism 2.37
13.2_amino acid metabolism.degradation 2.27
16.1.4_secondary metabolism.isoprenoids.carotenoids 2.40
16.8_secondary metabolism.flavonoids -2.04
20.1.7_stress.biotic.PR-proteins -2.14
20.2_stress.abiotic 0.0
20.2.99_stress.abiotic.unspecified 0.0
21_Redox 2.37
26_misc. -2.00
26.21_misc.protease inhibitor/seed storage/lipid transfer protein (LTP) family protein 0.0
27_RNA -2.34
27.3_RNA.regulation of transcription -2.18
27.3.11_RNA.regulation of transcription.C2H2 zinc finger family -2.15
29_protein 4.10
29.2_protein.synthesis 3.51
29.2.1_protein.synthesis.ribosomal protein 2.32
29.2.1.1_protein.sysnthesis.ribosomal protein.prokaryotic 0.0
29.2.1.1.1_protein.synthesis.ribosomal protein.prokaryotic.chloroplast 0.0
29.3_protein.targeting 2.41
29.4_protein.postranslational modification 0.0
29.4.1_protein.postranslational modification.kinase 0.0
29.4.1.57_protein.postranslational modification.kinase.receptor like cytoplasmatic kinase VII 0.0
29.5_protein.degradation 2.41
29.5.3_protein.degradation.cysteine protease 2.36
30.2_signalling.receptor kinases -4.11
30.2.11_signalling.receptor kinases.leucine rich repeat XI -2.14
30.2.17_signalling.receptor kinases.DUF 26 -2.14
31.1_cell.organisation 2.23
35_not assigned -5.63
35.2_not assigned.unknown -5.87
-3
-2
-1
0
1
2
3

overlapping functional classes.

The remaining 19 functional classes, which encompassed a total of 362 genes (Figure 3.4, Table 3.2, Appendix 1), were considered to be important components in the gene network regulating malate levels in developing fruit. Given its known role in acidity, *Ma1* was added to the list, making the total number of genes to be further analyzed at 363. Based on their putative functions, it could be postulated that: (i) several pathways might be critical for the variation of malate levels in developing fruit, including the malate and pyruvate interconversion reaction (8.2.10), photosynthesis (1.1.1.2 and 1.1.2.2), mitochondrial electron transport (9) and amino acid degradation (13.2), and (ii) genes in functional classes C2H2 zinc finger transcription factors (27.3.11, 14 genes), protein posttranslational modification (29.4.1.57, 27 genes), and signaling receptor kinases (30.2.11 and 30.2.17, 33 genes) were likely essential for transcriptional regulation (Table 3.2, Appendix 1).

K-means clustering of the 363 genes using software MeV 4.9 (Saeed et al. 2006)allowed grouping them into six clusters of 52, 132, 66, 54, 32 and 28 genes (Figure 3.5, Table 3.2, Appendix 1). The 132 genes in Cluster 2 were down-regulated in high malate fruit while the rest 231 in the other five clusters were all up-regulated in the high malate fruit (Figure 3.5).

3.4.5 Co-expression gene networks regulating malate concentrations

Using the network Inferring tool of software Cytoscape 3.1 , co-expression gene networks were constructed with 294 of 363 genes under Pearson correlation coefficient r>0.94, which included a major network of 286 members (Figure 3.6A) and 4 mini networks of 8 members (Figure 3.6B). Notably, *Ma1* was not present in these networks due to the threshold r>0.94. Analyzing the networks with a

Table 3.2. The number and K-mean cluster of genes in MapMan bins co-enriched or co-suppressed in varying malate groups or samples.

MapMan bin code	MapMan bin description	No. of Genes	Percent (%)	K-means cluster					
				I	II	III	IV	V	VI*
1.1.1.2	PS.lightreaction.photosystem II.PSII polypeptide subunits	29	8.0	22	2	5			
1.1.2.2	PS.lightreaction.photosystem I.PSI polypeptide subunits	18	5.0	17		1			
8.2.10	TCA / org transformation.other organic acid transformatons.malic	9	2.5		9				
9	Mitochondrial electron transport/ATP synthesis.NADH-DH.localisation not clear	9	2.5		8	1			
10.6	Cell wall.degradation	18	5.0		1	2	7	5	3
13.2	Amino acid metabolism.degradation	11	3.0	1	10				
16.1.4	Secondary metabolism.isoprenoids.carotenoids	8	2.2		8				
16.8	Secondary metabolism.flavonoids	30	8.3		5	2	20	2	1
20.1.7	Stress.biotic.PR-proteins	11	3.0	1	2	2	4	1	1
20.2.99	Stress.abiotic.unspecified	28	7.7		4	11	7	1	5
26.21	Misc.protease inhibitor/seed storage/lipid transfer protein (LTP) family protein	10	2.8	1	1	3	2	2	1
27.3.11	RNA.regulation of transcription.C2H2 zinc finger family	14	3.9	1	3	3	2	1	4
29.2.1.1.1	Protein.synthesis.ribosomal protein.prokaryotic.chloroplast	19	5.2	3	7	9			
29.3	Protein.targeting.nucleus	30	8.3	1	20	7	1	1	
29.4.1.57	Protein.postranslational modification.kinase.receptor like cytoplasmatic kinase VII	27	7.5		9	4	3	5	6
29.5.3	Protein.degradation.cysteine protease	19	5.2		13	3	1	1	1
30.2.11	Signalling.receptor kinases.leucine rich repeat XI	23	6.4	2	2	5	4	6	4
30.2.17	Signalling.receptor kinases.DUF 26	10	2.8	1	1	4	3	1	
31.1	Cell.organisation	39	10.8	2	27	4		5	1
Sum		362		52	132	66	54	31	27
%		100	100	14.4	36.5	18.2	14.9	8.6	7.5

* without *Ma1*

Figure 3.5 Expression of K-means clusters of the 363 genes or transcripts in MapMan functional classes that were significantly co-enriched or co-suppressed differentially in the fruit groups of high, mid and low malate. The vertical axis shows the square root transformation of RPKM values. Note the expression of genes in Cluster #2 that was negatively correlated with malate concentrations, and *Ma1* was included in Cluster #6.

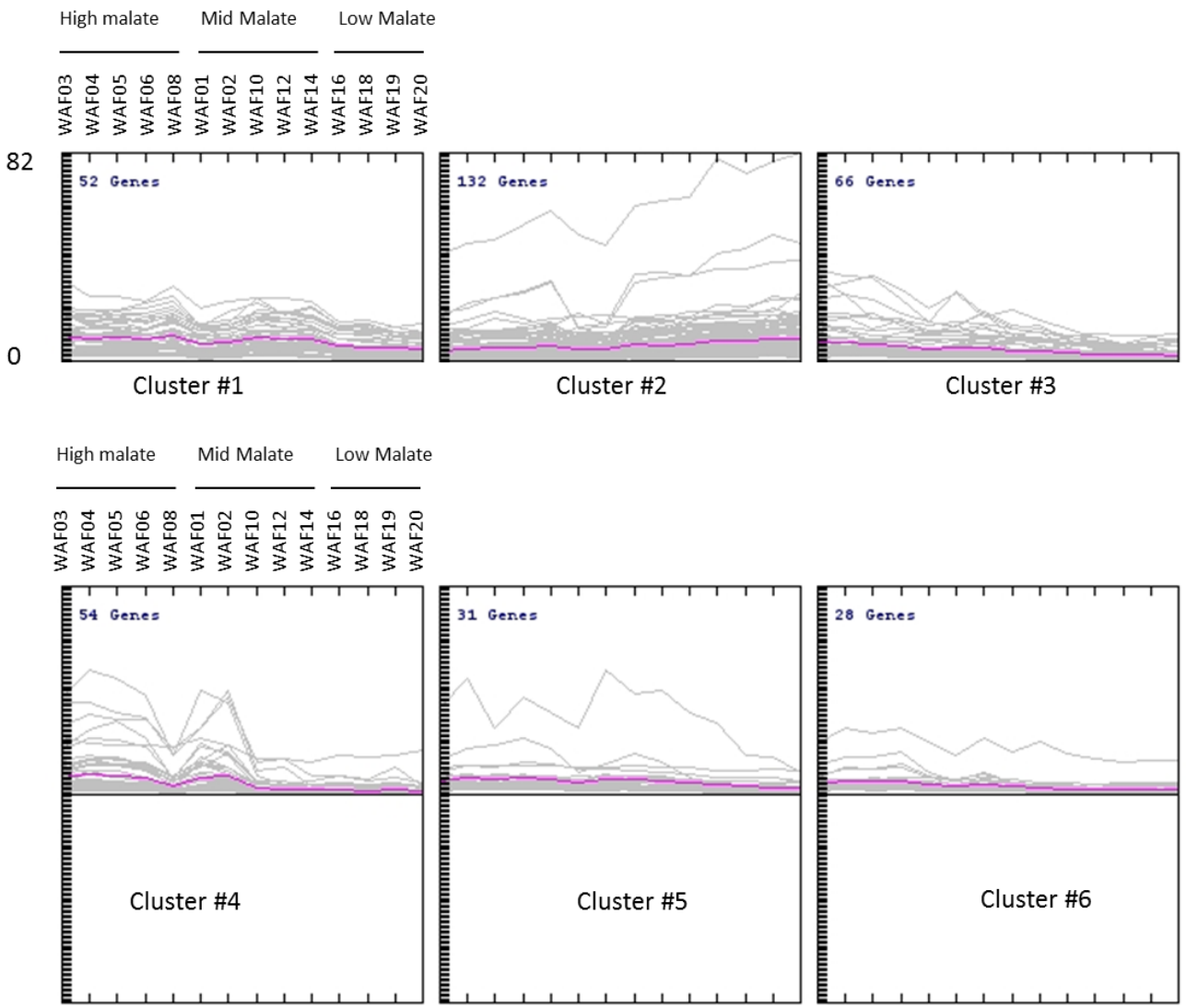
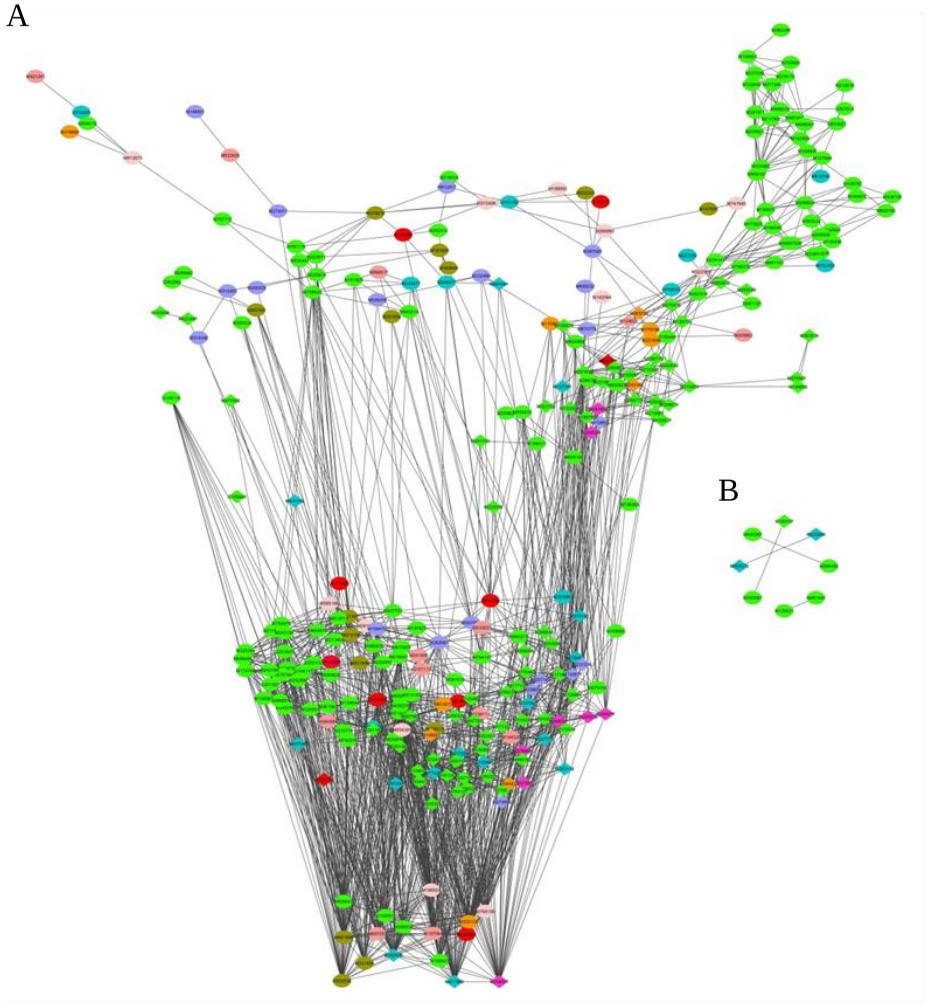
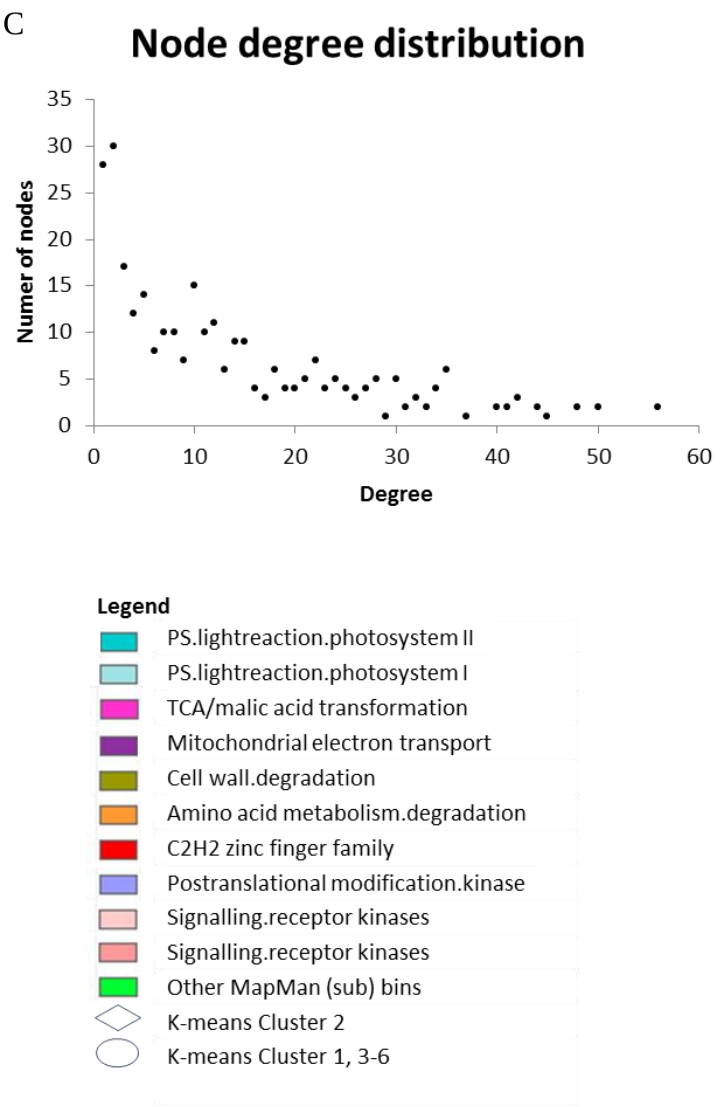

High malate
Mid Malate
Low Malate
WAF03
WAF04
WAF05
WAF06
WAF08
WAF01
WAF02
WAF10
WAF12
WAF14
WAF16
WAF18
WAF19
WAF20
82
0
52 Genes
132 Genes
66 Genes
Cluster #1
Cluster #2
Cluster #3
High malate
Mid Malate
Low Malate
WAF03
WAF04
WAF05
WAF06
WAF08
WAF01
WAF02
WAF10
WAF12
WAF14
WAF16
WAF18
WAF19
WAF20
54 Genes
31 Genes
28 Genes
Cluster #4
Cluster #5
Cluster #6

Figure 3.6 A graphic representation of the co-expression gene networks consisted of 294 of the 363 genes (shown in Figure 3.5) regulating malate levels in developing fruit. The networks were constructed using the Cytoscape network inferring tool (the Pearson r>0.94) and analyzed with NetworkAnalyzer (Assenov et al. 2008). Node color and shape keys are indicated in the legend. Please note that the diamond and oval shapes represent for K-means cluster 2 (negatively correlated with malate levels, see also Figure 3.5) and the remainder 5 clusters (positively correlated with malate levels), respectively. **A**. A major co-expression gene network of 286 members. The bottom section shows the 16 genes of the highest degrees at the network core, and mid-section shows the 121 immediate neighbors of the 16 genes at the core, and the top shows the rest (149 genes).**B**. Four mini networks of eight members. **C**. Degrees of nodes in the major and mini networks.


A
B


C
Node degree distribution
Numer of nodes
Degree
0
5
10
15
20
25
30
35
0
10
20
30
40
50
60
Legend
PS.lightreaction.photosystem II
PS.lightreaction.photosystem I
TCA/malic acid transformation
Mitochondrial electron transport
Cell wall.degradation
Amino acid metabolism.degradation
C2H2 zinc finger family
Postranslational modification.kinase
Signalling.receptor kinases
Signalling.receptor kinases
Other MapMan (sub) bins
K-means Cluster 2
K-means Cluster 1, 3-6

Cytoscape plugin NetwrokAnalyzer v2.7 (Assenov et al. 2008) showed that there were 16 nodes (genes) of the highest degrees (the number of neighbors) of 40-56 in the major network (Figure 3.6C, Table 3.3) while 75 nodes were of the lowest degrees with 28 of degree 1, 30 of degree 2 and 17 of degree 3 (Fiure.3.6A-C). The remainder (majority) 203 had degrees of 4-37.

The 16 genes of the highest degrees were considered as the core of the major gene network, and they collectively had 121 immediate neighbors (Figure 3.7). Revealing the identity of the 16 genes (Table 3.3) suggested that there were four (M186555, M798156, M897253 and M157044) putatively for encoding leucine rich repeat (LRR) receptor kinases in signaling, three (M252536, M321839 and M491898) for cell wall degradation related enzymes, one (M246502) for a C2H2 zinc finger transcription factor, one (M258977) for a NADP malic enzyme (ME), and seven for others. These results largely reinforced the two postulations above, i.e. (i) the pathways of the malate and pyruvate interconversion (e.g. by the NADP ME), photosynthesis, mitochondrial electron transport and amino acid degradation were essential for the variation of malate levels in developing fruit; and (ii) the co-expression gene network not only involved symplastic signal transduction (e.g. mediated by LRR receptor kinases), transcription regulation (e.g. by the C2H2 zinc finger transcription factor) and post-translational modification, but also the role of apoplast (e.g. through the cell wall degradation enzymes).

3.4.6 qRT-PCR confirmation of gene expression

To confirm if the RPKM values truly reflect the expression levels, three genes *Ma1* (M252114), M196894 and M127123 were analyzed with qRT-PCR assay (Figure 3.8). The relative expression of these three genes in qRT-PCR were highly correlated with

Table 3.3 List of genes of the highest node degrees in the core of the co-expression network

Feature ID	MapMan Bincode	Putative function	Chromosomal location	Degree
M258977	8.2.10	NADP-malic enzyme	Chr01_11311642_11314074-_MDC010772.65	44
M252536	10.6.1	Glycosyl hydrolase, protein endoglucanase 1 precursor, putative	Chr17_8305354_8308999-_MDC016922.37	42
M321839	10.6.2	1-4-beta-mannan endohydrolase, putative	Chr04_17599145_17601066-_MDC018286.199	40
M491898	10.6.3	Pectate lyase family protein	Chr13_7209618_7211439-_MDC009663.119	48
M295908	20.2.99	Pollen Ole e 1 allergen and extensin family protein	Chr09_4585826_4587577+_MDC012046.326	42
G100001	20.2.99	Pollen Ole e 1 allergen and extensin family protein	Chr09_4996315..4998730_MDC021673.67	41
M835003	26.21	Protease inhibitor/seed storage/lipid transfer protein (LTP) family protein	Chr14_27933836_27934129+_MDC020266.309	42
M246502	27.3.11	Zinc finger (C2H2 type) family protein	Chr11_7014798_7015913+_MDC029184.30	50
M192427	29.3.4.1	ER lumen protein retaining receptor, putative	Chr07_26304595_26306870+_MDC009045.262	44
M332125	29.3.4.99	SEC14 cytosolic factor family protein / phosphoglyceride transfer family protein	Chr12_25250564_25252501+_MDC013159.462	56
M798156	30.2.11	Leucine rich repeat (LRR) receptor kinase	Chr15_13028580_13031710+_MDC014159.224	50
M897253	30.2.11	Leucine rich repeat (LRR) receptor kinase	Chr12_21126633_21129746-_MDC012729.141	45
M186555	30.2.11	Leucine rich repeat (LRR) receptor kinase	Chr02_5517287_5518621+_MDC011209.233	41
M157044	30.2.17	Leucine-rich repeat family protein / protein kinase family protein	Chr15_24576085_24579174-_MDC000636.612	40
M382436	31.1	Plastid-lipid associated protein PAP/fibrillin family-like protein	Chr04_10225653_10239011-_MDC022173.275	56
M477969	31.1	Plastid-lipid associated protein PAP/fibrillin family-like protein	Chr06_2998593_3000895+_MDC000741.237	48

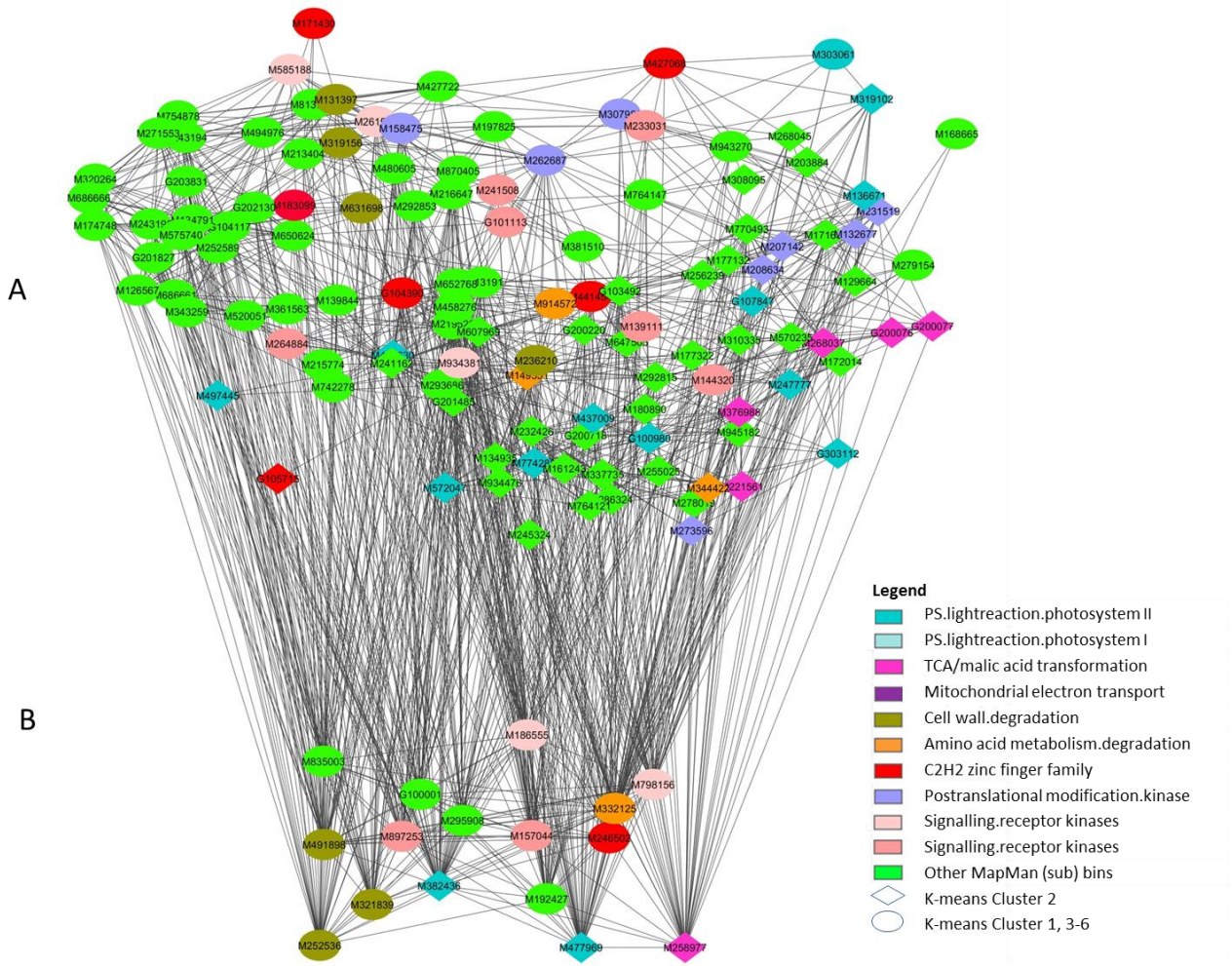


Figure 3.7 A detailed view of the 16 core members (B) and their immediate 121 neighbors (A) in the co-expression network. The node color and shape keys are the same as noted in Figure 3.6.

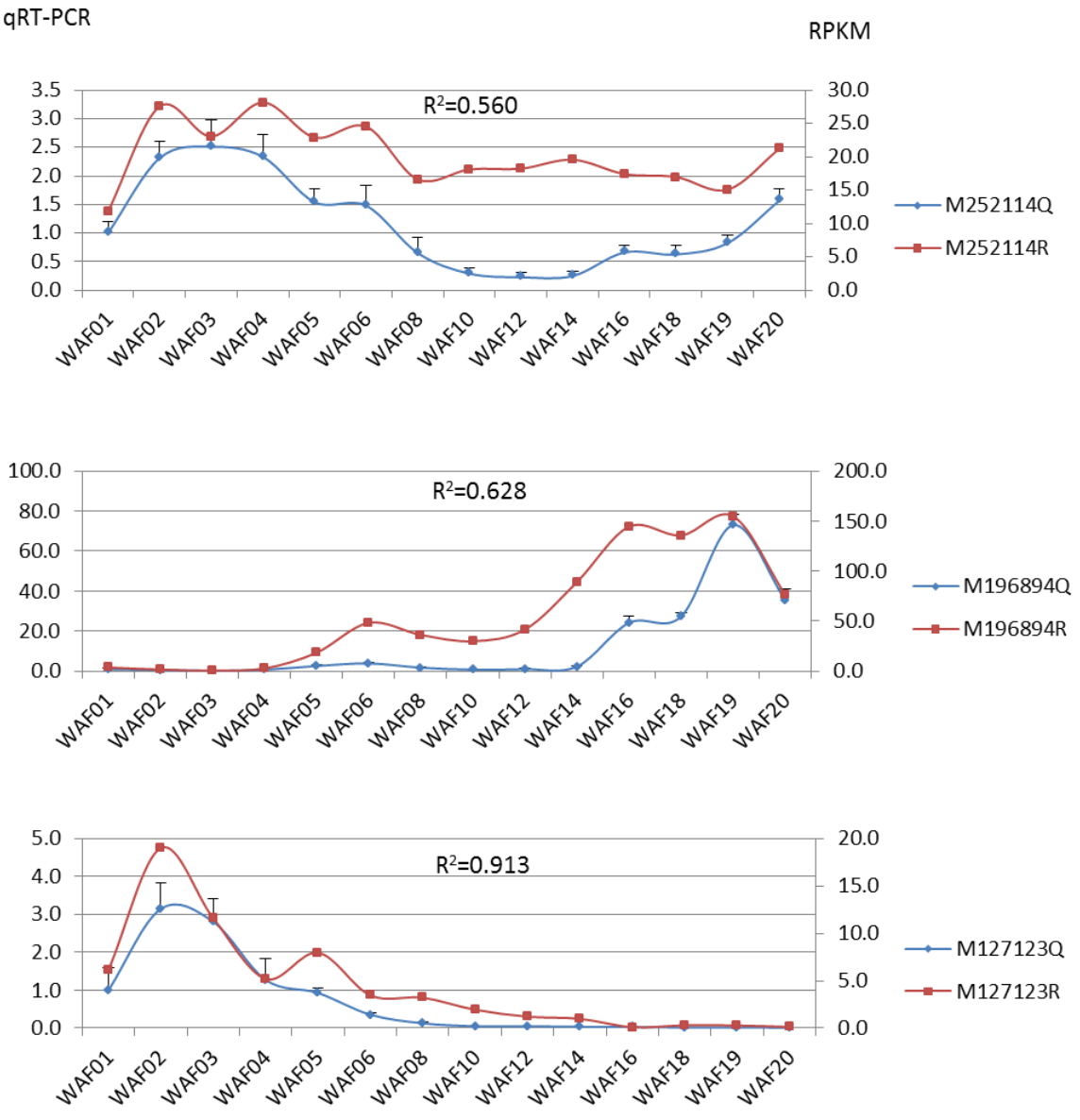


Figure 3.8 Confirmation of gene expression of three selected genes using qRT-PCR. The normalized expression of target genes relative to a control gene (actin) in qRT-PCR was shown in blue (against the primary vertical axis), and their corresponding RPKM values from RNA-seq were in red (against the secondary vertical axis). The efficient of determination (R^2) was shown accordingly.

their RPKM values with *Ma1* of r^2=0.560, M196894 of r^2=0.628, and M127123 of r^2=0.913, which all exceeded the critical value (r^2=0.471) at the significance level of $P_{0.01}$.

3.5 Discussion

3.5.1 Identification of genes associated with the variation of malate levels in developing fruit

Changes of malate levels in developing fruit has long been reported and are often characterized by a peak in a period from WAF04-WAF08 (Beruter 2004; Hulme and Wooltorton 1957; Ulrich 1970; Zhang et al. 2010). In this study, we found a similar developmental pattern in malate levels in fruit of Golden Delicious, i.e. it was increased rapidly from WAF02 through WAF5 and then progressively decreased through harvest (WAF20). The malate concentrations varied from 5.15 mg/g fw (WAF19) to 14.52±0.48 mg/g fw (WAF05) in the 14 time points measured and could be grouped into high, mid and low malate (Figure 3.1). We hypothesized this developmental change in fruit malate levels is regulated by a dynamic gene network that may involve *Ma1*, the major gene considered to control apple fruit acidity at maturity (Bai et al. 2012; Khan et al. 2013). To shed light on the gene network, we took a transcriptomics approach by analyzing a set of RNA-seq data obtained from the 14 time points in developing fruit using the improved apple reference transcriptome (Bai et al. 2014). We found that there were 39,789, 40,714, and 37,213 genes expressed in the high, mid, and low malate groups, respectively (Figure 3.2A), which collectively corresponded to 43,496 genes expressed. Of these expressed genes, 3,066 were expressed not only differentially (P_{FDR}<0.05) between the high and low malate

groups but also in significant ($P<0.05$) correlation with malate concentrations and/or the expression of *Ma1* in the 14 time points. Confirmation the expression through qRT-PCR on three of them (M252114 (*Ma1*), M196894 and M127123) indicated that the expression measured in RNA-seq was significantly ($R^2=0.56$-0.91, $P<0.01$) correlated with that in qRT-PCR. Based on these data, we predict that a majority, if not all, of the genes that are responsible for malate variations in developing fruit are included in the 3,066 genes.

3.5.2 Pathways putatively important in determining malate levels

To extract relevant biological information from this large set of 3,066 genes, we elected to use the MapMan annotation of functional classes that were designed for plants (Thimm et al. 2004) rather than the general GO (Gene Ontology) terms. Based on the comparative study on the two ontologies MapMan and GO (Langmead et al. 2009), and the objective of this study, it appeared that MapMan based annotation was more appropriate. The 3,066 genes were assigned into 648 functional classes or MapMan (sub-) bins, of which 1,949 genes (63.6%) were of a putative function while 1117 (36.4%) of unknown function (Figure 3.3).

The PageMan tool (Usadel et al. 2006) has been used widely on plants to condense large data set by identifying enriched pathways or MapMan functional classes (Gowik et al. 2011; Langmead et al. 2009; Secco et al. 2013; Usadel et al. 2009). Using this tool, we found that 19 of the 648 MapMan functional classes encompassing the 3,066 genes were significantly ($P_{FDR}<0.05$) co-enriched or co-suppressed in a malate dependent manner (Figure 3.4, Table 3.2). The 19 functional classes comprised 363 (362 plus *Ma1*) genes, and K-means clustering of these genes revealed that expression of 132 (36.5%) of them was negatively correlated with malate

levels while the expression of the rest 231 positively (Figure 3.5, Table 3.2). The 19 functional classes included known primary metabolic pathways for malate synthesis and degradation, such as the malate and pyruvate interconversion reaction and photosynthesis, as well as those that have indirect effects on malate, such as mitochondrial electron transport and amino acid metabolism.

The malate and pyruvate interconversion reaction is one of the key steps underlying the current model of potential metabolic pathways involving malate in fruit (Etienne et al. 2013; Sweetman et al. 2009). The interconversion reaction occurs in cytosol and in chloroplast and is more favored from malate to pyruvate catalyzed by NADP dependent malic enzymes (NADP-ME). The identification of sub-bin 8.2.10 of nine NADP-ME encoding genes as one of the 19 functional classes strongly suggested the role of the malate and pyruvate interconversion reaction in malate levels. Since they were expressed significantly higher only in low malate group and were all fallen into K-means Cluster 2, the nine NADP-ME encoding genes appeared to negatively affect malate levels in fruit, i.e. higher expression leads to low malate levels. Based on the sequence similarity annotations of the nine NADP-MEs (Appendix 1), six were similar to AtNADP ME1, two (including M258977 at the core of the gene network) to AtNADP ME3, and one to AtNADP ME4. Since enzymes AtNADP-ME2 and AtNADP-ME3 were cytosolic and AtNADP ME4 was plastidic (Wheeler et al. 2005), the negative correlation was likely a reflection of the cytosolic reaction rather the one in plastids. Indeed, enzyme activities of a characterized cytosolic NADP-ME (DQ280492), which is nearly identical to M192078, has been shown to negatively contribute to malate accumulation in apple fruit (Yao et al. 2009), although the difference in the expression of M192078 was insignificant between the high and low malate groups in this study. Moreover, a recent study on the plastidic NADP-ME that was knocked-out by RNAi in ripening tomato fruit showed that malic acid was

reduced (Osorio et al. 2013), suggesting that disrupting the interconversion between pyruvate and malate in plastid would reduce rather than increase the malic acid levels in fruit.

Developing apple fruit are capable of photosynthesis although the CO_2 assimilated is primarily from the mitochondrial respiration (Blanke and Lenz 1989). In this pathway, oxaloacetate is produced through the action of phosphoenolpyruvate carboxylase (PEPC) and then malate by NADP malate dehydrogenase (NADP-MDH). The expression of the two functional classes (sub-bins) 1.1.1.2 of 29 genes associated with photosystems I (PSI) and 1.1.2.2 of 18 from PSII were significantly higher in high- and mid- malate groups while normal in low acid group (Figure 3.4). K-means clustering of these two sub-bins indicated that 45 of the 47 genes falling into Clusters 1 and 3 that were positively correlated with malate levels (Table 3.2). These data suggested a more active photosynthesis process in relatively young fruit may have facilitated the biosynthesis of malate, contributing to high malate levels in the high-malate group (WAF3-WAF8).

Mitochondrial electron transport is one of the major respiratory pathways in which the reducing equivalents (e.g. NADH) generated from the TCA cycle are used to drive the synthesis of ATP (FJ and TK 1998). While generating ATP, the complex I of electron transport chain also makes NAD^+ available from NADH for the TCA cycle where the recycling of NAD^+ is required. Therefore, the TCA cycle and the electron transport chain are in fact an integrated process, largely explaining the putative role of mitochondrial electron transport in fruit malate levels. There were nine genes in the mitochondrial electron transport functional class (bin 9), with three for complex I, two for complex III and four for ATP synthases (Table 3.2, Appendix 1). The nine genes were expressed at significantly higher levels in low-malate group but normal in high- and mid-malate groups (Figure 3.4), and eight of them were in K-means Cluster II

(Table 3.2, Figure 3.5). These data seemed to suggest that high expression of genes in the electron transport chain was correlated with low malate levels. This is consistent with the finding that malate and citrate increased in the leaves of the CMSII mutant in *Nicotiana sylvestris* that lacks functional complex I (Gutierres et al. 1997; Noctor et al. 2004).

The 11 genes identified in the functional class amino acid metabolism-degradation (sub-bin 13.2) were expressed at normal levels in high- and mid-malate groups but at significantly higher levels in the low malate groups (Figure 3.5). Of these genes, four were for degradation of the aspartate family amino acids (asparagine, lysine, threonine, and methionine), five for the aromatic amino acids including tyrosine (4) and tryptophan (1), and two for branched chain amino acids (Leucine or isoleucine or valine). Although these amino acids have their unique degradation pathways, the pathways could converge into intermediates acetyl CoA, fumarate, pyruvate, oxaloacetate, and succinyl-CoA (Nelson et al. 2008). Obviously, all of the intermediates can enter the TCA cycle, the known major pathway for malate metabolism, accounting for the involvement of these amino acid degradation genes.

3.5.3 Co-expression gene networks regulating malate concentrations

Using the 363 genes in the 19 MapMan functional classes that were co-enriched or co-suppressed depending upon malate groups, we inferred a major co-expression gene network of 286 genes. There were 16 genes at the core of the network based on the nodes of the highest degrees (Figures 3.6, 3.7, Table 3.2), which were assumed to have more important roles in the network. The major gene *Ma1* for fruit acidity was not present in the network due to the threshold $r>0.94$ used in the network construction. However, it is still possible that the variation in malate concentration actually involved

Ma1. This could be evidenced from the fact that 97 of the 286 genes in the network were initially identified through their significant correlations with *Ma1* although 88 of these genes also showed high correlation with malate concentrations. In grape, a *Ma1* like gene *VvALMT9* has also been demonstrated to mediate both malate and tartrate accumulation in developing berries (Angeli et al. 2013).

Members of the ALMT1 protein family are known for their critical role in aluminum tolerance in plants (Collins et al. 2008; Hoekenga et al. 2006; Sasaki et al. 2004), where the efflux organic anions especially malate from roots is a common mechanism. Under the current transcriptional regulation model of aluminum tolerance in Arabidopsis (Delhaize et al. 2012), *sensitive to proton rhizotoxicity1* (*STOP1*), a C2H2 zinc finger transcriptional factor, controls the expression of *AtALMT1*, *multidrug and toxic compound extrusion 1* (*AtMATE1*), *aluminum sensitive 3* (*AtALS3*), and other proton and Al^{3+} responsive genes (Delhaize et al. 2012). It was suggested that protein kinases might activate STOP1 by phosphorylation and the activated STOP1 would then bind to the promoters of the targeted genes to initiate transcription. However, the sensor elements interacting with Al^{3+} and/or protons were unknown. A model comprising similar elements was also proposed to account for the Al^{3+} tolerance in rice (Delhaize et al. 2012). The counterpart of *STOP1* is *Al^{3+} resistance transcription factor 1* (*ART1*), another C2H2 transcription factor, and ART1 had been proven to regulate the expression of at least 31 genes related to detoxification mechanisms (MR and MR 2004). The 19 functional classes that were expressed in a malate dependent manner in developing fruit included important elements in the aluminum tolerance models, such as C2H2 transcriptional factors (14 genes in sub-bin 27.3.11), protein kinase for post translational modification (27 genes in sub -bin 29.4.1.57) and leucine rich repeat (LRR) receptor kinases for signaling (23 genes in sub-bin 30.2.11 and 10 genes in sub-bin 30.2.17). Interestingly, G105811, a

new gene identified from contig MDC016907.364 in the improved reference transcriptome shared the highest sequence similarities with STOP1 in Arabidopsis and is the only one annotated as an ortholog of STOP1 by Mercator. It would be of interest to investigate if any of these genes would directly or indirectly interact with *Ma1*.

The identification of the 18 genes (sub-bin 10.6) associated with cell wall degradation suggested that regulation of malate levels in developing fruit might also involve the roles of apoplast (e.g. pectin). There were two lines of evidence in the literature supporting the possible role of apoplast in fruit malate levels: 1) Pectin content and degree of methylation in root-apex was relevant for Al-tolerance where malate efflux is an important mechanism (H et al. 2001; JK 2002; SJ and SH 1998; Wehr et al. 2003). 2) The shift of fruit phloem unloading from symplastic to apoplastic pathway prior to maturity was coupled with sharper decline of malate in vacuole than in apoplast in ripening grape berries (Keller and Shrestha 2014; Zhang et al. 2006). In other fruits, such as cucumber (Hu et al. 2011) and tomato (Ruan et al. 1995), similar fruit phloem unloading shift was reported as well. In apple (Golden Delicious) developing fruit, phloem unloading was also observed to be via apoplast (Zhang et al. 2004).

The roles of genes in other seven functional classes related to secondary metabolism (16.1.4 and 16.8), stress (20.1.7 and20.2.99), protein synthesis (29.2.1.1.1), protein targeting (29.3), protein degradation (29.5.3), cell organization (31.1) and miscellaneous (26.21) will not be discussed due to the complexity of their functions and the limited scope of the study. Obviously, whether or not and how any of these 363 genes would play a role in regulating fruit malate levels are questions that can only be answered by functional studies in the future. An important direction would be to investigate how gene *Ma1* might interact with the genes putatively encoding

receptor kinases (30.2.11 and 30.2.17), protein kinases (29.4.1.57) and C2H2 transcription factors (27.3.11).

In conclusion, through analysis of RNA-seq data derived from Golden Delicious fruit at 14 developmental stages, we identified 19 functional classes that were co-enriched or co-suppressed in a malate dependent manner, which encompassed 363 genes. The putative function of these genes suggests that several pathways might be critical for the variation of malate in developing fruit, including the malate and pyruvate interconversion reaction, photosynthesis, mitochondrial electron transport and amino acid degradation. The identification of functional classes of C2H2 zinc finger transcription factors, protein posttranslational modification, and signaling receptor kinases suggests that they are likely essential for transcriptional regulation in the major co-expression gene network regulating malate levels in developing fruit, which were inferred to comprise 286 of 363 genes. Despite the complexity of the gene network, this study has provided transcriptomics clues relevant for better understanding the change of malate levels in developing apple fruit.

References

<SORBITOL DEHYDROGENASE EXPRESSION IN APPLE FRUIT.pdf>

Angeli A, Baetz U, Francisco R, Zhang J, Chaves M, Regalado A (2013) The vacuolar channel VvALMT9 mediates malate and tartrate accumulation in berries of Vitis vinifera. Planta 238:283-291

Assenov Y, Ramírez F, Schelhorn S-E, Lengauer T, Albrecht M (2008) Computing topological parameters of biological networks. Bioinformatics 24:282-284

Bai Y, Dougherty L, Li M, Fazio G, Cheng L, Xu K (2012) A natural mutation-led truncation in one of the two aluminum-activated malate transporter-like genes at the Ma locus is associated with low fruit acidity in apple. Molecular Genetics and Genomics 287: 663-678

Bai Y, Dougherty L, Xu K (2014) Towards an improved apple reference transcriptome using RNA-seq. Molecular Genetics and Genomics published online:1-12

Barbier-Brygoo H, De Angeli A, Filleur S, Frachisse J-M, Gambale F, Thomine S, Wege S (2011) Anion Channels/Transporters in Plants: From Molecular Bases to Regulatory Networks. Annu Rev Plant Biol 62:25-51

Benjamini Y, Hochberg Y (1995) Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. J R Stat Soc Ser B-Methodol 57:289-300

Beruter J (2004) Carbohydrate metabolism in two apple genotypes that differ in malate accumulation. Journal of Plant Physiology 161:1011-1029

Blanke MM, Lenz F (1989) Fruit photosynthesis. Plant, Cell & Environment 12:31-46

Collins NC, Shirley NJ, Saeed M, Pallotta M, Gustafson JP (2008) An ALMT1 gene cluster controlling aluminum tolerance at the Alt4 locus of rye (Secale cereale L.). Genetics 179:669-682

Delhaize E, Ma JF, Ryan PR (2012) Transcriptional regulation of aluminium tolerance genes. Trends in Plant Science doi: 10.1016/j.tplants.2012.02.008

Emmerlich V, Linka N, Reinhold T, Hurth MA, Traub M, Martinoia E, Neuhaus HE (2003) The plant homolog to the human sodium/dicarboxylic cotransporter is the vacuolar malate carrier. Proceedings of the National Academy of Sciences of the United States of America 100:11122-11126

ES L, LM L, B B, C N, MC Z, Gen-ome IH, Consortium S (2001) Initial sequencing and analysis of the human genome. Nature 409:860

Etienne A, Génard M, Lobit P, Mbeguié-A-Mbéguié D, Bugaud C (2013) What controls fleshy fruit acidity? A review of malate and citrate accumulation in fruit cells. Journal of Experimental Botany 1451-1469

Ferguson IB (1984) Calcium in plant senescence and fruit ripening. Plant, Cell & Environment 7:477-489

FJ T, TK F (1998) Characterization of two glutamate decarboxylase cDNA clones from Arabidopsis. Plant Physiol 117:1411

Gasic K, Hernandez A, Korban SS (2004) RNA extraction from different apple tissues rich in polyphenols and polysaccharides for cDNA library construction. Plant Molecular Biology Reporter 22:437-438

Gowik U, Bräutigam A, Weber KL, Weber APM, Westhoff P (2011) Evolution of C4 Photosynthesis in the Genus Flaveria: How Many and Which Genes Does It Take to Make C4? The Plant Cell Online 23:2087-2105

Gutierres S, Sabar M, Lelandais C, Chetrit P, Diolez P, Degand H, Boutry M, Vedel F, de Kouchkovsky Y, De Paepe R (1997) Lack of mitochondrial and nuclear-encoded subunits of complex I and alteration of the respiratory chain in Nicotiana sylvestris mitochondrial deletion mutants. Proceedings of the National Academy of Sciences 94:3436-3441

H Z, M B, R B, D H, A C (2001) Global analysis of protein activities using proteome chips. Science 293:2101

Hoekenga OA, Maron LG, Pineros MA, Cancado GMA, Shaff J, Kobayashi Y, Ryan PR, Dong B, Delhaize E, Sasaki T, Matsumoto H, Yamamoto Y, Koyama H, Kochian LV (2006) AtALMT1, which encodes a malate transporter, is identified as one of several genes critical for aluminum tolerance in Arabidopsis. Proceedings of the National Academy of Sciences of the United States of America 103:9738-9743

Hu L, Sun H, Li R, Zhang L, Wang S, Sui X, Zhang Z (2011) Phloem unloading follows an extensive apoplasmic pathway in cucumber (Cucumis sativus L.) fruit from anthesis to marketable maturing stage. Plant, Cell & Environment 34:1835-1848

Hulme AC, Wooltorton LSC (1957) The organic acid metabolism of apple fruits: changes in individual acids during growth on the tree. Journal of the Science of Food and Agriculture 8:117-122

JK Z (2002) Salt and drought stress signal transduction in plants. Annu Rev Plant Biol 53:247

Keller M, Shrestha P (2014) Solute accumulation differs in the vacuoles and apoplast of ripening grape berries. Planta 239:633-642

Kenis K, Keulemans J, Davey M (2008) Identification and stability of QTLs for fruit quality traits in apple. Tree Genetics & Genomes 4:647-661

Khan S, Beekwilder J, Schaart J, Mumm R, Soriano J, Jacobsen E, Schouten H (2013) Differences in acidity of apples are probably mainly caused by a malic acid transporter gene on LG16. Tree Genetics & Genomes 9:475-487

Kovermann P, Meyer S, Hortensteiner S, Picco C, Scholz-Starke J, Ravera S, Lee Y, Martinoia E (2007) The Arabidopsis vacuolar malate channel is a member of the ALMT family. Plant Journal 52:1169-1180

Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10:R25

Liebhard R, Kellerhals M, Pfammatter W, Jertmini M, Gessler C (2003) Mapping quantitative physiological traits in apple (Malus x domestica Borkh.). Plant Molecular Biology 52:511-526

Lohse M, Nagel A, Herter T, May P, Schroda M, Zrenner R, Tohge T, Fernie AR, Stitt M, Usadel B (2014) Mercator: a fast and simple web server for genome scale functional annotation of plant sequence data. Plant, Cell & Environment Published online:n/a-n/a

Maliepaard C, Alston FH, van Arkel G, Brown LM, Chevreau E, Dunemann F, Evans KM, Gardiner S, Guilford P, van Heusden AW, Janse J, Laurens F, Lynn JR, Manganaris AG, den Nijs APM, Periam N, Rikkerink E, Roche P, Ryder C, Sansavini S, Schmidt H, Tartarini S, Verhaegh JJ, Vrielink-van Ginkel M, King GJ (1998) Aligning male and female linkage maps of apple (Malus pumila Mill.) using multi-allelic markers. Theoretical and Applied Genetics 97:60-73

Meyer S, Scholz-Starke J, De Angeli A, Kovermann P, Burla B, Gambale F, Martinoia E (2011) Malate transport by the vacuolar AtALMT6 channel in guard cells is subject to multiple regulation. Plant Journal 67:247-257

Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nature Methods 5:621-628

MR R, MR K (2004) Oxidative stress-induced calcium signaling in Arabidopsis thaliana. Plant Physiol 135:1471

Nelson DL, Lehninger AL, Cox MM (2008) Pathways of Amino Acid Degradation. Lehninger principles of biochemistry, 5th edn. W.H. Freeman, New York, pp 687-706

Noctor G, Dutilleul C, De Paepe R, Foyer CH (2004) Use of mitochondrial electron transport mutants to evaluate the effects of redox state on photosynthesis, stress tolerance and the integration of carbon/nitrogen metabolism. Journal of Experimental Botany 55:49-57

Osorio S, Vallarino JG, Szecowka M, Ufaz S, Tzin V, Angelovici R, Galili G, Fernie AR (2013) Alteration of the Interconversion of Pyruvate and Malate in the Plastid or Cytosol of Ripening Tomato Fruit Invokes Diverse Consequences on Sugar But Similar Effects on Cellular Organic Acid, Metabolism, and Transitory Starch Accumulation. Plant Physiology 161:628-643

Pfaffl MW (2001) A new mathematical model for relative quantification in real-time RT-PCR. Nucleic Acids Research 29:e45

Ruan Y, Mate C, Patrick J, Brady C (1995) Non-Destructive Collection of Apoplast Fluid From Developing Tomato Fruit Using a Pressure Dehydration Procedure. Functional Plant Biology 22:761-769

Saeed A, Bhagabati N, Braisted J, Liang W, Sharov V, Howe E, Li J, Thiagarajan M, White J, Quackenbush J (2006) TM4 microarray software suite. Methods Enzymol 411:134-193

Sasaki T, Yamamoto Y, Ezaki B, Katsuhara M, Ahn SJ, Ryan PR, Delhaize E, Matsumoto H (2004) A wheat gene encoding an aluminum-activated malate transporter. Plant Journal 37:645-653

Schumacher K, Krebs M (2010) The V-ATPase: small cargo, large effects. Current Opinion in Plant Biology 13:724-730

Secco D, Jabnoune M, Walker H, Shou H, Wu P, Poirier Y, Whelan J (2013) Spatio-Temporal Transcript Profiling of Rice Roots and Shoots in Response to Phosphate Starvation and Recovery. The Plant Cell Online 25:4285-4304

SJ Y, SH O (1998) Cloning and characterization of a tobacco cDNA encoding calcium/calmodulin-dependent glutamate decarboxylase. Mol Cells 8:125

Sweetman C, Deluc LG, Cramer GR, Ford CM, Soole KL (2009) Regulation of malate metabolism in grape berry and other developing fruits. Phytochemistry 70:1329-1344

Thimm O, Blasing O, Gibon Y, Nagel A, Meyer S, Kruger P, Selbig J, Muller LA, Rhee SY, Stitt M (2004) MAPMAN: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. Plant Journal 37:914-939

Ulrich R (1970) Organic acids. In: Hulme A (ed) The biochemistry of fruit and their products. Academic Press, London and New York, pp 89-118

Usadel B, Nagel A, Steinhauser D, Gibon Y, Blasing O, Redestig H, Sreenivasulu N, Krall L, Hannah M, Poree F, Fernie A, Stitt M (2006) PageMan: An interactive ontology tool to generate, display, and annotate overview graphs for profiling experiments. BMC Bioinformatics 7:535

Usadel B, Poree F, Nagel A, Lohse M, Czedik-Eysenberg A, Stitt M (2009) A guide to using MapMan to visualize and compare Omics data in plants: a case study in the crop species, Maize. Plant Cell and Environment 32:1211-1229

Wang A, Xu K (2012) Characterization of Two Orthologs of REVERSION-TO-ETHYLENE SENSITIVITY1 in Apple. Journal of Molecular Biology Research 2:24-41

Wehr JB, Menzies NW, Blamey FPC (2003) Model studies on the role of citrate, malate and pectin esterification on the enzymatic degradation of Al- and Ca-pectate gels: possible implications for Al-tolerance. Plant Physiology and Biochemistry 41:1007-1010

Wheeler MCG, Tronconi MA, Drincovich MF, Andreo CS, Flügge U-I, Maurino VG (2005) A Comprehensive Analysis of the NADP-Malic Enzyme Gene Family of Arabidopsis. Plant Physiology 139:39-51

Wilcoxon F (1945) Individual comparisons by ranking methods. Biometrics 1:80-83

Xu K, Wang A, Brown S (2012) Genetic characterization of the Ma locus with pH and titratable acidity in apple. Molecular Breeding 30:899-912

Yao Y-X, Li M, Zhai H, You C-X, Hao Y-J (2011) Isolation and characterization of an apple cytosolic malate dehydrogenase gene reveal its function in malate synthesis. Journal of Plant Physiology 168:474-480

Yao YX, Li M, Liu Z, You CX, Wang DM, Zhai H, Hao YJ (2009) Molecular cloning of three malic acid related genes MdPEPC, MdVHA-A, MdcyME and their expression analysis in apple fruits. Scientia Horticulturae 122:404-408

Zhang LY, Peng YB, Pelleschi-Travier S, Fan Y, Lu YF, Lu YM, Gao XP, Shen YY, Delrot S, Zhang DP (2004) Evidence for apoplasmic phloem unloading in developing apple fruit. Plant Physiology 135:574-586

Zhang XY, Wang XL, Wang XF, Xia GH, Pan QH, Fan RC, Wu FQ, Yu XC, Zhang DP (2006) A shift of phloem unloading from symplasmic to apoplasmic pathway is involved in developmental onset of ripening in grape berry. Plant Physiology 142:220-232

Zhang YZ, Li PM, Cheng LL (2010) Developmental changes of carbohydrates, organic acids, amino acids, and phenolic compounds in 'Honeycrisp' apple flesh. Food Chemistry 123:1013-1018

CHAPTER 4

UNCOVERING THE *MA1* ASSOCIATED CO-EXPRESSION GENE NETWORK GOVERNING APPLE FRUIT ACIDITY

4.1 Introduction

Apple fruit acidity refers to the sensory intensity of tardiness or sour taste of fruit flesh tissues. The stronger the sourness taste, the higher the fruit acidity levels. Chemically, fruit acidity can be quantified by measuring fruit juice pH and/or titratable acidity (TA). It has been shown that several organic acids are present in mature apple fruit, including malic acid, quinic acid, citric acid and others, but malic acid accounts for more than 90% of the total (Hulme and Wooltorton 1957; Yamaki 1984; Zhang et al. 2010), thereby largely determining fruit acidity. For dessert apples, the acceptable range of fruit acidity was estimated of 3.0-10.0 mg/ml. This makes fruit acidity an essential quality component not only determining the fate of existing varieties, but also prompting routine evaluations of acidity levels in apple breeding as genotypes of fruit acidity beyond the acceptable range can make up 25-50% in breeding populations. Comprehensibly, fruit acidity has long been an important subject area of investigations in apple genetics.

The current view of genetic control of apple fruit acidity is that the trait is primarily governed by the major gene or QTL on chromosome 16, called *Malic acid* (*Ma*) alongside a significant QTL on chromosome 8 and a few other QTLs of relatively smaller effects (Kenis et al. 2008; Kumar et al. 2012; Liebhard et al. 2003; Maliepaard et al. 1998; Zhang et al. 2012a). With regard to the allelic interactions at the major QTL *Ma*, high acid allele (*Ma*) is nearly completely dominant over the low

acid allele (*ma*). Genotypes *MaMa* and *Mama* set fruits of normal to high acidity while genotype *mama* of extremely low acidity with little or no commercial value. To focus on the most important genetic factor, we (Bai et al. 2012) and others (Khan et al. 2013) have recently isolated the major QTL *Ma.* The common findings were that the *Ma* locus harbors two new members of the *Aluminum-activated Malate Transporter1* (*ALMT1*) gene family, called *Ma1* and *Ma2*. The studies further found that *Ma1* was expressed in significant positive correlation with fruit acidity levels while the expression of *Ma2* was barely detectable in both high and low acid fruit, suggesting that it was gene *Ma1* rather than *Ma2* that was the very gene underlying *Ma* (Bai et al. 2012; Khan et al. 2013). In addtion, a detailed analysis of the allele specific DNA sequences of *Ma1* indicated that a single base mutation that would stop the protein translation process prematurely was almost completely associated with low acidity in fruit studied, suggesting that the low acidity is caused by the malfunction of the MA1 protein due to the deduced truncation at the C-terminus (Bai et al. 2012).

These latest findings by revealing the *Ma* locus have markedly increased our understanding on fruit acidity, but many remain to be learned. The immediate questions to be answered include: Why is *Ma1* transcribed at higher levels in genotypes *MaMa* and *Mama* than in genotype *mama* if the stop codon leading mutation was the cause for low acidity? How does the stop codon leading mutation in *Ma1* affect the landscape of transcriptomes in genotype *mama* in relation to *MaMa* and *Mama*? If the *Ma1* is a central regulator in an integrated gene network governing fruit acidity, what would be the other possible members? The development of mRNA sequencing (RNA-seq) technology that unlocks the power of high throughput next generation sequencing (NGS) has provided an ideal means to address these questions. Since its inception (Mortazavi et al. 2008; Wilhelm and Landry 2009), RNA-seq has been rapidly adapted in transcriptomics studies in plants such as *Arabidopsis* (Lister et

al. 2008), grape (Zenoni et al. 2010), maize (Li et al. 2010) and rice (MJ et al. 1998). In apple, RNA-seq based studies have recently been reported as well (Krost et al. 2013; Krost et al. 2012; M et al. 1999; T et al. 1995; Xia et al. 2012; Zhang et al. 2012b). To resolve the low coverage issue of the current version of apple reference transcriptome, we improved it with RNA-seq reads from fruit of Golden Delicious, the source of the reference genome.

The objectives of this study are to address the questions mentioned above. To do so, we first sequenced 30 RNA-seq libraries representing transcriptomes of mature fruit of varying acidity levels from ten apple varieties of genotypes *MaMa*, and *Mama* and *mama*, and then analyzed the RNA-seq data using the improved apple reference transcriptome.

4.2. Materials and methods

4.2.1 Plant materials and fruit acidity quantification

Ten apple varieties of known genotypes at the *Ma* locus were chosen, including four of *mama* - Britegold, Sweet Delicious, Novosibirski Sweet and PI323617, four of *Mama* - Fuji (Red Sport Type 2), Rome Beauty Law, Cox's Orange Pippin and Jonathan and two of *MaMa* - Empire and Granny Smith. Genotypes *Mama* and *MaMa* were jointly designated *Ma*__ to represent genotypes that had at least one functional allele of *Ma1*. The trees were budded onto rootstock P21 and grown in a research orchard of Cornell University, Geneva, NY, USA. Fruit of three replicates per variety and 8-10 fruit per replicate were harvested at maturity from 2-3 trees in fall, 2012. Each fruit was cross-sectioned into two halves: one half were used for maturity evaluation and fruit juice extraction and the other half were sliced and frozen in liquid

nitrogen for RNA isolation and for quantitation of malate and other metabolites. The evaluation of fruit maturity and fruit acidity was conducted as previously described (Xu et al. 2012). Fruit of Cornell Starch Index 4.0-6.0 (Blanpied and Silsby 1992) were considered matured and only matured fruit (with the core removed) were used for analysis. Fruit acidity was measured with both pH by a pH meter (Accumet AB15, Fisher Scientific, PA, USA) and titratable acidity (TA) by an autotitrator (Metrohm 848 Titrino Plus with 869 compact sample changer, Metrohm, Herisau, Switzerland).

Fruit organic acids were extracted and derivatized following a protocol described previously (Lisec et al. 2006) with minor modifications, which was addressed in detail in Bai et al. (2014a). Briefly, metabolites were extracted from 100 mg homogenized apple fruit tissue using 1.4 ml of 75% methanol with 600 ppm ribitol added as internal standard and then dried under vacuum, which were followed by two steps of derivatization reactions. The derivatized metabolites were analyzed with an Agilent 7890A GC/5795C MS (Agilent Technology, Palo Alto, CA, USA) with the same configurations and settings as described in Bai et al. (2014a). Metabolites were identified by comparing fragmentation patterns with those in mass spectral libraries and quantified based on standard curves for each metabolite and the internal standard ribitol. Statistical analyses, such as ANOVA and the Student's t test were performed in JMP Pro10 (SAS, Cary, NC).

4.2.2 RNA isolation and strand specific RNA-seq library construction and sequencing

RNA isolation, and RNA-seq library construction were performed as previously described in Bai et al. (2014b). Briefly, total RNA was isolated from 3g of ground fruit tissue and treated with DNase I (amplification grade, Invitrogen/Life Technologies, Carlsbad, CA). As described in Bai et al. (2014b), NEBNext Poly(A)

mRNA Magnetic Isolation Module and NEBNext Ultra Directional RNA Library Prep Kit for Illumina (New England Biolabs, Ipswich, MA) were used for mRNA isolation and stand specific RNA-seq library construction. The mRNA was isolated from 5 ug of total RNA and was fragmented at 94°C for 10min. First strand cDNA was reverse transcribed from the fragmented mRNA with dATP mix and second strand cDNA was synthesized from the first strand cDNA with dUTP mix. The resulting double strand cDNA was end-repaired, adaptor-ligated, size-selected, and followed by USER enzyme digestion of the second strand cDNA and PCR enrichment in 14-16 cycles. The libraries were multiplexed in equal amount for single-end 100 base sequencing in three lanes of HiSeq 2000 (Illumina, San Diego, CA) at the Cornell University Biotechnology Resource Center (Ithaca, NY). The ten apple varieties were sequenced in three biological replicates with one replicate per lane.

4.2.3 RNA-seq data analysis

RNA sequencing generated 30 sequence FASTQ files with a total of 615.6 million reads (Table 4.1). The raw reads were fed into Bowtie (Langmead et al. 2009) to remove bar codes and adapters, followed by aligning to rRNA database downloaded from http://www.arb-silva.de allowing up to three mismatches (Quast et al. 2013). Data analyses were performed using CLC Genomics Workbench (CLC GW) v6.5 (CLCBio, Cambridge, Massachusetts). The rRNA depleted reads were imported into CLC GW to trim low quality reads and/or bases using the quality limit of 0.05 and the ambiguous limit of 1. The resultant clean and high quality reads were mapped against the improved apple reference transcriptome (Bai et al. 2014b) using the minimum similarity fraction 0f 0.98, the minimum length fraction of 0.8 and the maximum number of hits of 10. For convenience, hereafter the novel reference

Table 4.1 Overview of RNA-seq reads mapping

Sample Name	Raw Reads	Clean and high quality reads	Total mapped reads		Uniquely mapped reads	
			Count	Rate(%)	Count	Rate(%)
B_rep1	19,988,850	18,090,212	13,224,684	73.1	11,050,718	61.1
B_rep2	18,390,359	16,665,545	12,390,438	74.3	10,357,109	62.1
B_rep3	13,774,683	12,937,762	9,713,920	75.1	8,129,087	62.8
C_rep1	10,999,347	5,858,386	4,196,304	71.6	3,440,510	58.7
C_rep2	9,961,580	8,287,353	6,258,449	75.5	5,221,241	63.0
C_rep3	13,850,408	9,205,224	6,814,163	74.0	5,664,150	61.5
E_rep1	25,330,875	19,672,562	14,664,581	74.5	12,031,039	61.2
E_rep2	17,735,691	14,002,497	10,577,742	75.5	8,673,332	61.9
E_rep3	17,996,572	12,998,952	9,758,083	75.1	7,877,473	60.6
F_rep1	20,616,175	19,255,968	14,553,591	75.6	12,193,928	63.3
F_rep2	27,314,760	23,879,202	17,883,722	74.9	14,956,353	62.6
F_rep3	16,001,675	14,160,346	10,636,482	75.1	8,906,461	62.9
G_rep1*	18,464,198	3,487,013	1,922,103	55.1	1,265,630	36.3
G_rep2	21,853,897	18,401,994	14,030,668	76.2	11,637,676	63.2
G_rep3	23,204,576	19,472,936	14,828,356	76.1	12,318,014	63.3
J_rep1	20,946,545	15,194,756	11,379,660	74.9	9,319,118	61.3
J_rep2	17,072,375	12,447,289	9,421,763	75.7	7,775,122	62.5
J_rep3	30,600,975	20,954,078	15,579,561	74.4	12,778,713	61.0
N_rep1	33,446,343	26,090,450	18,438,899	70.7	15,081,832	57.8
N_rep2	26,028,574	20,575,728	14,794,890	71.9	12,124,196	58.9
N_rep3	21,201,256	19,016,910	13,688,174	72.0	11,317,824	59.5
P_rep1	16,246,019	10,560,623	7,719,453	73.1	6,425,807	60.8
P_rep2	12,162,206	8,528,412	6,325,021	74.2	5,268,151	61.8
P_rep3	12,564,429	9,994,648	7,379,126	73.8	6,153,560	61.6
R_rep1	27,347,497	24,529,890	18,744,681	76.4	15,691,355	64.0
R_rep2	23,186,447	19,602,928	15,012,272	76.6	12,496,053	63.7
R_rep3	24,034,622	21,525,886	16,449,310	76.4	13,787,899	64.1
S_rep1	25,150,325	20,191,803	15,196,181	75.3	12,613,314	62.5
S_rep2	23,260,887	17,619,738	13,302,850	75.5	10,967,333	62.2
S_rep3	26,868,746	15,048,868	10,994,109	73.1	8,877,105	59.0
Total	615,600,892	478,257,959	355,879,236	/	294,400,103	/
Mean	20,520,030	15,941,932	11,862,641	73.9	9,813,337	60.8
St Dev	5,808,998	5,561,988	4,199,348	3.8	3,515,793	4.8

*The mapping rate of G_rep1 is too low and this sample is not used in the downstream RNA-seq analysis.

transcripts will also be referred to as ‘genes’ in text.

Gene expression levels were calculated and normalized by reads per kilobase of exon model per million mapped reads (RPKM) (Mortazavi et al. 2008). Genes of RPKM>0.3 were defined expressed according to a previous study (Kang et al. 2013).

4.2.4 Identification of genes expressed differentially between genotypes *mama* and *Ma__* and in correlation with *Ma1*

To identify differentially expressed genes between genotypes *mama* and *Ma__*, the RPKM data were subjected to the Baggerly’s test (Baggerly et al. 2003). The original *p*-values in the Baggerly’s test were adjusted for multiple testing using Benjamini-Hochberg correction to control the false discovery rate (Benjamini and Hochberg 1995). The cutoff for a gene expressed differentially was $P_{FDR}<0.05$. For the differentially expressed genes, the correlation coefficients (Pearson’s *r*) with *Ma1* expression levels were calculated in Microsoft Excel 2010. The significance of correlation was set to $P_{r>0.632}<0.05$.

4.2.5 Inferring of the *Ma1* associated co-expression gene network

The RPKM data of the selected genes were square root transformed and averaged for each apple variety and then analyzed in Cytoscape 3.1 (Saito et al. 2012). The networks were inferred using the basic correlation method with $|r|>0.9$ and analyzed using the Cytoscape plugin Network Analyzer (Assenov et al. 2008).

4.2.6 q RT-PCR Analysis

Q RT-PCR analyses were conducted with the same total RNA samples (after DNase I treatment) used for RNA-seq library construction. Two micrograms of total RNA were reverse transcribed using the Superscript III RT module (Invitrogen/Life technology, Carlsbad, CA). The resulting first strand cDNA was diluted by fivefold and used as template for q RT-PCR analysis on LightCycler 480 (Roche, Indianapolis, IN), where an apple actin gene (EB136338) served as reference. The primer sequences of the reference gene and the target ten genes were listed in Table 4.2. For each reaction, a final volume of 16 µl was used, containing 5 µl of the cDNA dilutions, 0.5 µM forward and reverse primers and 1× SYBR green master mix (Roche, Indianapolis, IN). A standard curve for each gene was generated by a four level serial dilution of the template cDNA (1/5, 1/50, 1/500 and 1/5000). The qPCR program includes an initial denaturation step of 10 min at 94°C, a 45-cycle amplification of 10 s at 94°C, 25 s at 55°C, and 25 s at 72°C, and a dissociation stage of 5 s at 95°C, 60 s at 60°C, and 15 s at 97°C.

Expression quantification and data analysis were performed by software LightCycler 480 v1.5 using the comparative cycle threshold method (Pfaffl 2001). Amplification efficiency (E) of each gene was calculated as E=10 [-1/slope] using the slope of its standard curve. Statistical tests such as ANOVA and the Tukey's HSD (honest significant difference) test were performed in JMP Pro10 (SAS, Cary, NC).

4.3. Results

4.3.1 Fruit metabolite profiling and acidity evaluation

Fruit metabolite profiling was conducted using GC-MS with three biological replicates in the ten apple varieties. A total of 19 metabolites were quantified, including 12

Table 4.2 List of primers used in qRT-PCR

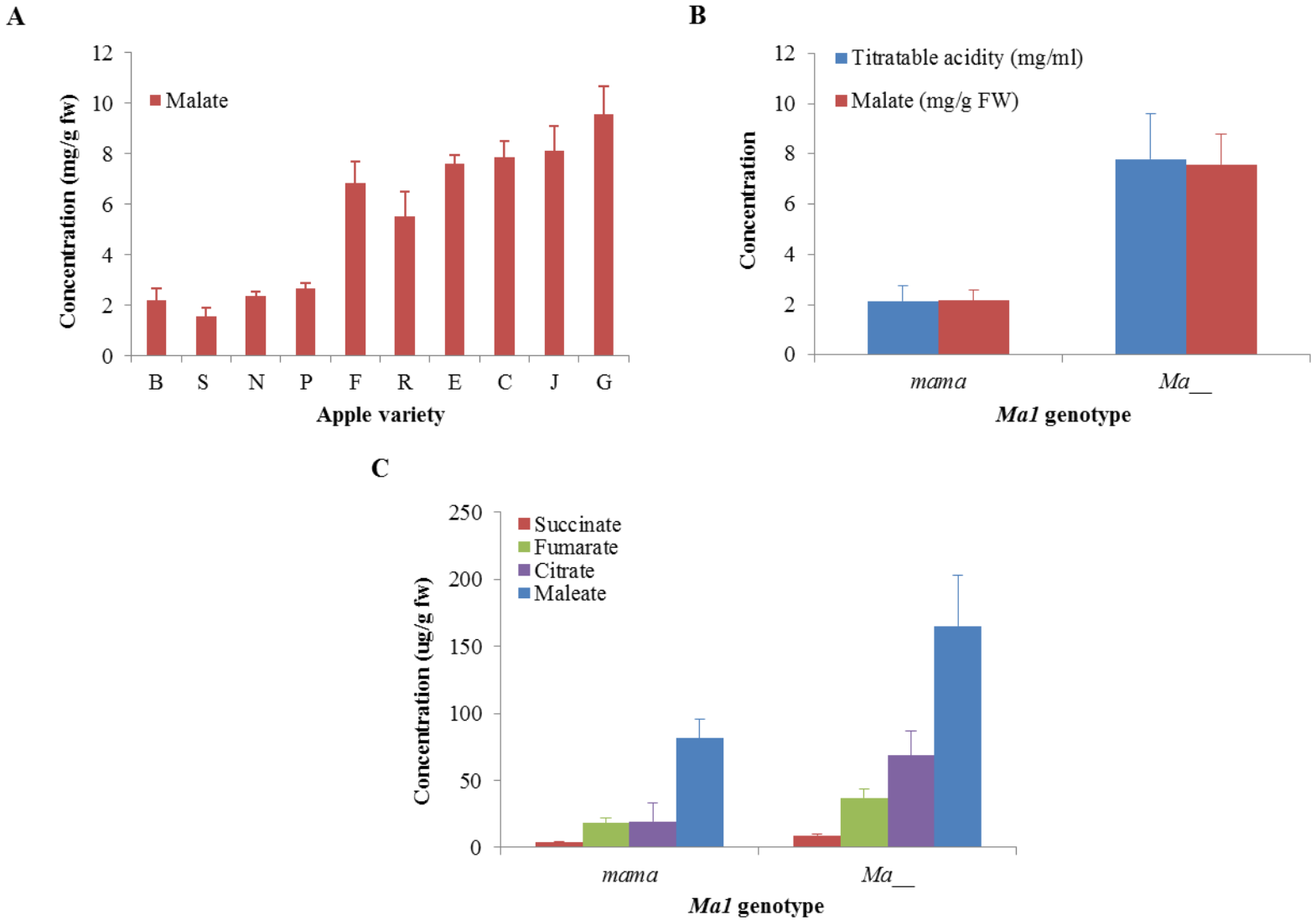
Primer ID	Primer Seq
2F_M_219042	CACCCAAACCTGATGAGAAGA
2R_M_219042	TGCTTCAACAGCTTCACTGG
3F_M_442350	GTGGTGAGTGAGAGGGTGGT
3R_M_442350	TTCTGCACTGTCCTATCCAAGA
4F_M_651862	ATTGACGATCAAGTGGAGCA
4R_M_651862	ATCTGGGACCGGACAACTC
6F_M_349063	TTGCGATCTCAATCACGAAA
6R_M_349063	CAAGAGTGAAAGAGACAAAGCTGA
11F_M_190273	GCACAAGATGGAATCCTGAAA
11R_M_190273	CAACCTTCTTCCTCCCTGAA
13F_M_132720	TCAGCTTGAGAGGGTGAATGT
13R_M_132720	TCACAATTCTCGGCCTCTTT
14F_M_815327	AATGAAGATCGTTGTGAAGGTG
14R_M_815327	GACGAGGTCGGCTTTACTTCT
15F_M_163222	TACCACCACTTTGCTCCACA
15R_M_163222	TCATTTCTCTCCCACGGATT
16F_M_196894	GCATCACGAAGAAGACGATG
16R_M_196894	TTCTTGCCGTGAATCAACAA
Ma1F_M_252114	GTACTCCGACTTGGGCTTCA
Ma1R_M_252114	ACATCTTTGAGCGGCACTTT
ActionF_EB136338	GGCTGGATTTGCTGGTGATG
ActionR_EB136338	TGCTCACTATGCCGTGCTCA

soluble sugars and seven organic acids. Among the 12 sugars, only sorbitol concentrations were significantly ($P<0.001$) higher in genotype *mama* (7.42±2.01 mg/g) than in genotype *Ma__* (5.37±1.89 mg/g), while other sugar levels, especially the high abundant fructose, sucrose and glucose, were consistent across the varieties. Among the seven organic acids - malate, dehydroascorbate, maleate, succinate, fumarate, citrate, and quinate, the most abundant acid was malate (89.7% of the total acidity), followed by quinate (6.2%) and maleate (2.2%). Malate concentrations were significantly ($P<0.001$) different across the ten apple varieties (Figure 4.1A) and more than three times significantly ($P<0.001$) higher in genotype *Ma__* than in *mama* (Figure 4.1B). Fruit titratable acidity (TA) was in very similar level as malate (Figure 4.1B). Specifically, the group-mean malate concentrations were 2.16±0.41 and 7.58±1.23 mg/g and the group-mean TA were 2.15±0.61 and 7.80±1.79 mg/ml in the *mama* and *Ma__* groups, respectively (Figure 4.1B). The concentrations of maleate, succinate, fumarate and citrate were also significantly different between the *mama* and *Ma__* groups ($P<0.001$ in t test) (Figure 4.1C).

4.3.2 Gene expression analysis in groups of genotypes *mama* and *Ma__*

Gene expression analysis was conducted using the improved apple reference transcriptome (Bai et al. 2014b). After removing low quality reads derived from rRNA, the total reads for RNA-seq mapping were 478.3 million. Overall the total mapped and uniquely mapped reads were 355.9 million (73.9%) and 294.4 million (60.8%) (Table 4.1). The mean mapped reads per sample were 11.9±4.2 million in total and 9.8±3.5 million in unique (Table 4.1). The RNA-seq sample of Granny Smith replicate I had high percentage of rRNA reads and low mapping rate and thus was excluded from the downstream analysis.

Figure 4.1 Organic acid concentrations in *Ma1* genotypes of *mama* and *Ma__*. Standard deviations were shown with the error bars. A. Malate concentration across the ten apple varieties. B. Malate concentration and titratable acidity in different *Ma1* genotypes. C. Succinate, fumarate and citrate and maleate concentrations in different *Ma1* genotypes.


A
Malate
Concentration (mg/g fw)
0
2
4
6
8
10
12
B S N P F R E C J G
Apple variety
B
Titratable acidity (mg/ml)
Malate (mg/g FW)
Concentration
mama
Ma__
Ma1 genotype
C
Succinate
Fumarate
Citrate
Maleate
Concentration (ug/g fw)
0
50
100
150
200
250
mama
Ma__
Ma1 genotype

Based on the cut off of RPKM>0.3, there were 50,951 genes expressed in at least one sample (Figure 4.2). Considering the mean expression levels of the two genotype groups, the total number of expressed genes (RPKM>0.3 in at least one group) was 41,465 (Figure 4.3A). In individual genotype group, there were 38,348 genes expressed in the *mama* group and 39,854 in the *Ma__* group with 36, 737 in common between the two (Figure 4.3A). Consistent with the group pattern of organic acids, the mean *Ma* 1 expression level in the *Ma__* group was significantly ($P<0.001$) higher than its expression in the *mama* group.

In terms of the total mapped reads of the expressed genes, there were 10.7±2.7 million reads per sample in the *mama* group and 11.1±3.1 million reads per sample in the *Ma__* group (Figure 4.3B). Regarding the uniquely mapped reads, there were 9.2±2.3 million and 9.6 ±2.7 million reads per sample in the *mama* and *Ma__* groups, respectively (Figure 4.3B).

4.3.3 Identification of genes expressed differentially between genotype groups *mama* and *Ma__* and in association with *Ma1*

To identify genes expressed in a genotype-dependent manner, the expression data were compared and subjected to the Baggerly's test (Baggerly et al. 2003). It showed that there were 716 genes expressed significantly ($P_{FDR}<0.05$) differentially between the *mama* and *Ma__* groups (Figure 4.2). To examine how the 716 differentially expressed genes were expressed in correlation with *Ma1*, we calculated their correlation coefficients (Pearson's *r*) and identified 303 significantly ($P_{r>0.632}<0.05$) correlated genes (Appendix 2). Since these 303 genes were not only expressed differentially between the two genotypes *mama* and *Ma__*, but also in correlation with *Ma1*, we concluded that they were relevant members of the *Ma1* regulated

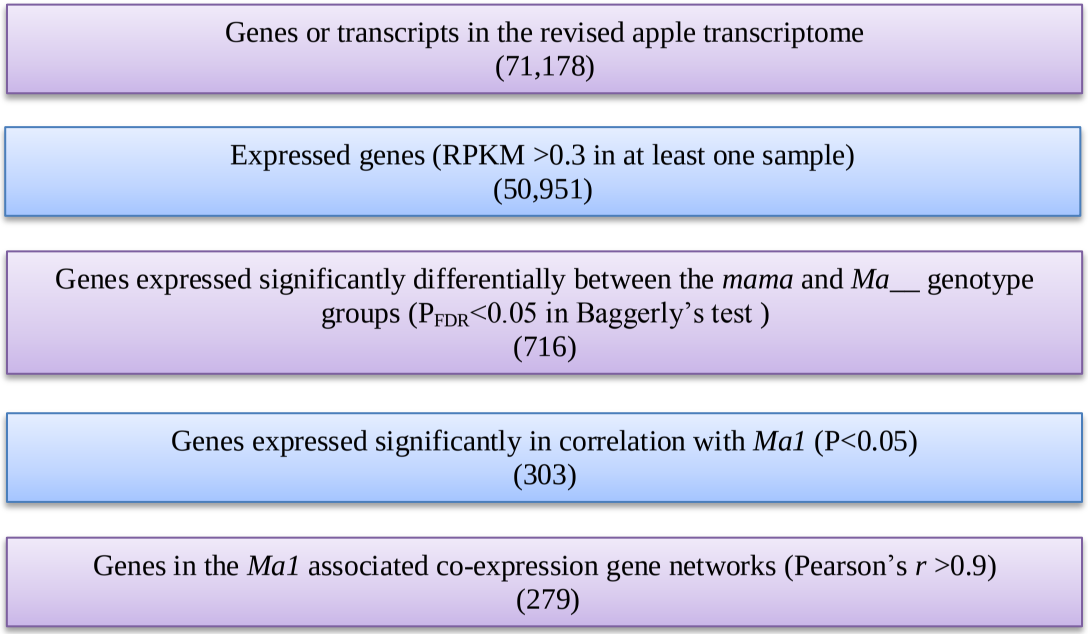


Figure 4.2 Workflow of RNA-seq data analysis using the improved apple reference transcriptome

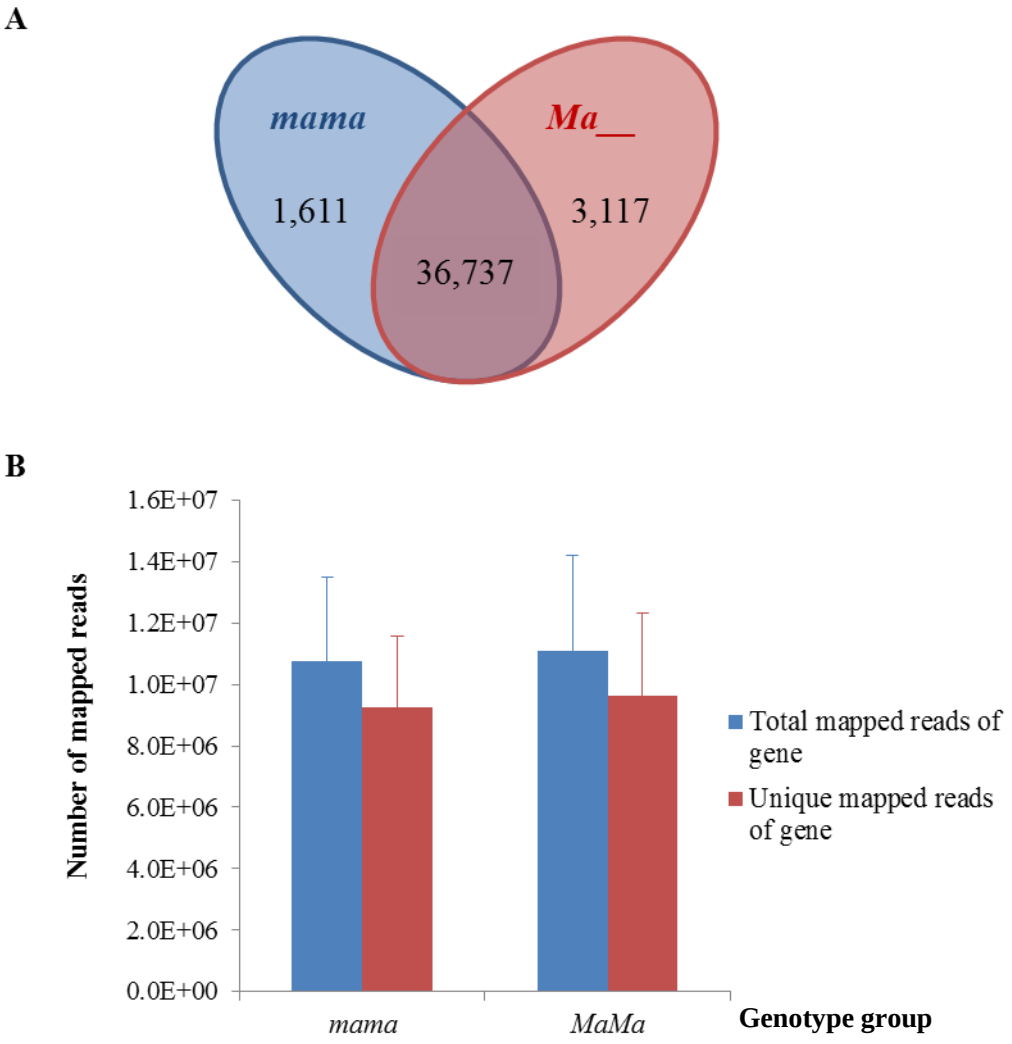


Figure 4.3 Overview of RNA-seq data analysis. **A.** Venn diagram representation of the number of genes expressed in genotype *mama* (blue) and *Ma__* (red) (RPKM>0.3). **B.** Mean number of total and unique mapped reads per sample in the groups of *Ma1* genotypes *mama* and *Ma__*. Standard deviations were shown with the error bars.

co-expression network controlling fruit acidity. Consistent with previous observations, the expression of *Ma1* was significantly correlated with fruit acidity in mature apple (with malate $P_{r=0.832}<0.05$ and with TA $P_{r=0.725}<0.05$).

MapMan gene ontology analysis showed that 187 (61.7%) of 303 genes were assigned to one or more MapMan functional bins while the remainder 116 (38.3%) were not assigned (Figure 4.4). Bins '29_protein' of 34 (11.2%) genes, '27_RNA' of 17 (5.6%), '20_stress biotic' of 16 (5.3%), '26_miscellaneous' of 16 (5.3%) and '30_signaling of 16 (5.3%) were the largest bins among the assigned whereas '2_major CHO metabolism' '4_glycolysis', '6_gluconeogenesis', '9_ Mitochondrial electron transport' and '18_ Co-factor and vitamin metabolism' were the smallest bins of only one gene (Figure 4.4). In the unassigned 116 genes in Bin 35, 42 (36.2%) were found of significant hits, but there were no gene ontology assigned in MapMan; and the rest 74 (63.8%) were unknown.

4.3.4 Co-expression gene network associated with *Ma1*

Using the Cytoscape Networking interfering tool (Cyni) and the Pearson's correlation threshold r>0.9, a major *Ma1* associated co-expression network of 264 nodes (genes) was constructed from the 303 genes (Figure 4.5A). Analyzing the networks with Cytoscape plugin NetworkAnalyzer v2.7 (Assenov et al. 2008) showed that there were 23 nodes of the highest degrees of 13-25 while 143 nodes were of the lowest degrees of 1-5(Figure 4.5C). The 23 genes of the highest degrees were considered to be of greater regulatory roles in the gene network governing fruit acidity, especially *Ma1* and those of putative roles in calcium and light signaling (M319170, M140330, M841118 and M307855), transcription regulation (M423596), protein post-translational modification (M651862 and M250124) and the TCA cycle (M682401),

Figure 4.4 Distribution in MapMan bins of the 303 genes which were expressed not only significantly (p<0.05) in correlation with *Ma1* expression, but also significantly (P_{FDR}<0.05) in difference between the *mama* and *Ma__* groups. The MapMan bins were coded with numbers 1-35, prefixing the bin name with a dash.

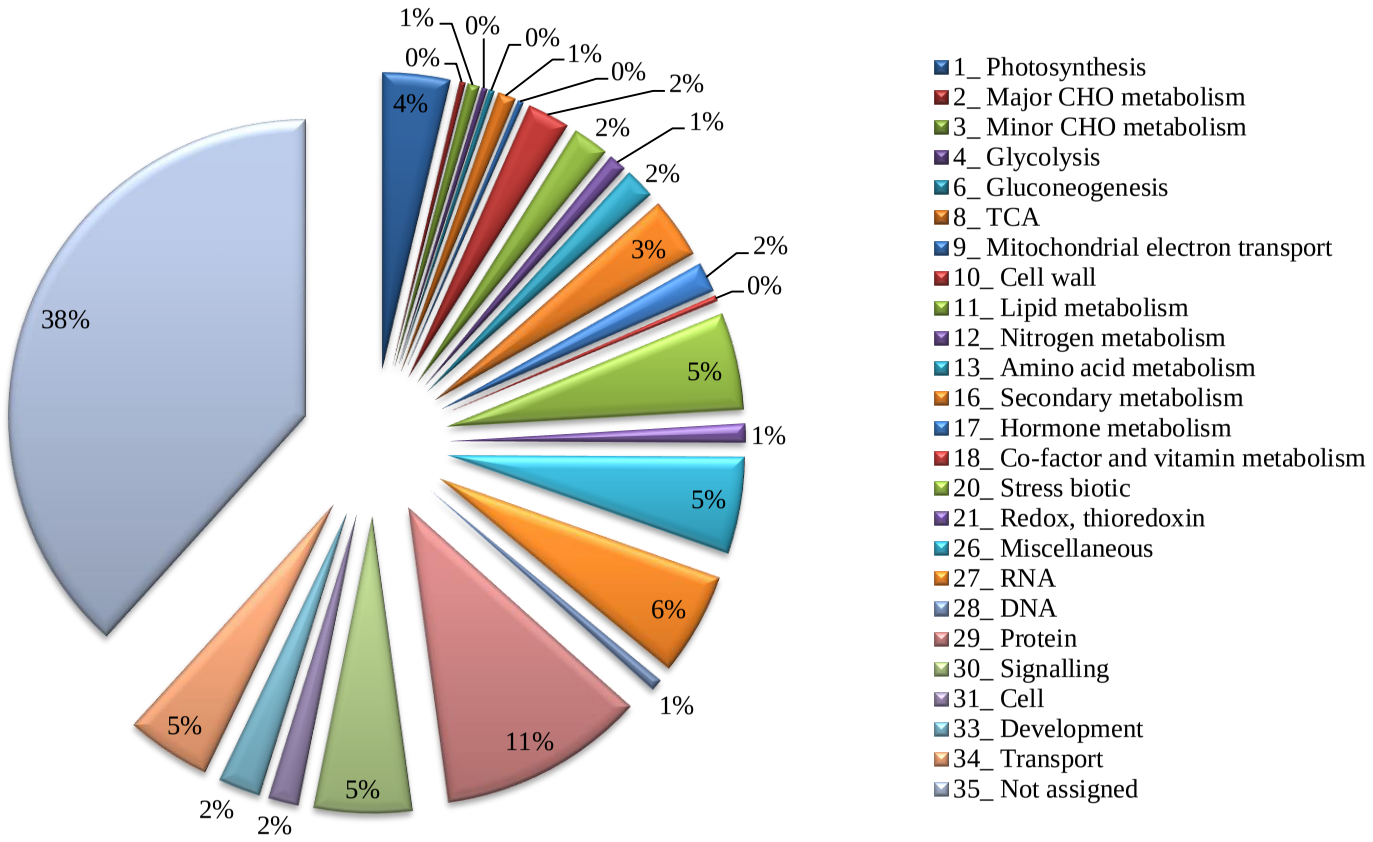
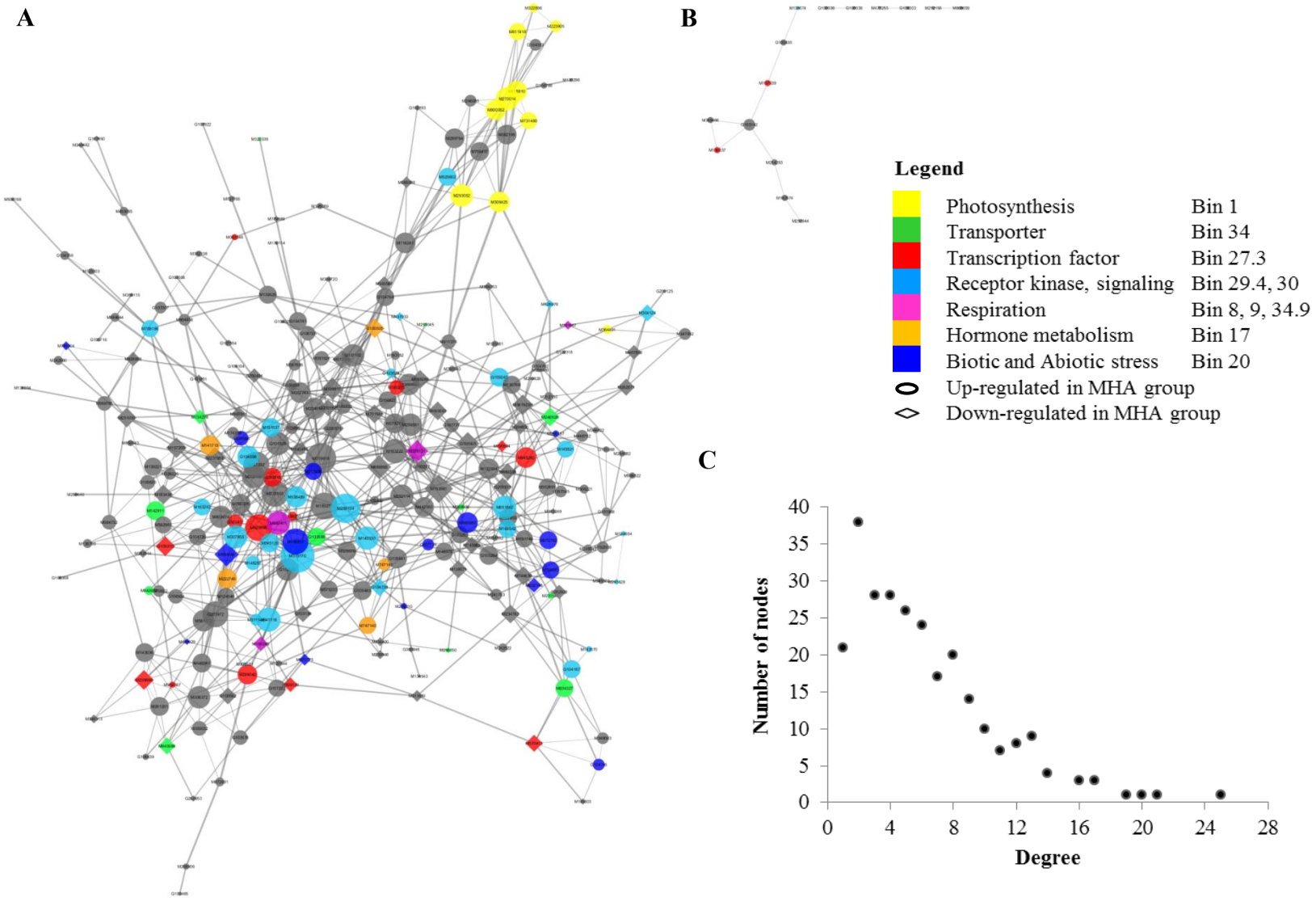

1%
0%
0%
0%
1%
0%
2%
4%
2%
1%
2%
3%
2%
0%
5%
1%
5%
6%
1%
11%
5%
2%
2%
5%
38%
1_ Photosynthesis
2_ Major CHO metabolism
3_ Minor CHO metabolism
4_ Glycolysis
6_ Gluconeogenesis
8_ TCA
9_ Mitochondrial electron transport
10_ Cell wall
11_ Lipid metabolism
12_ Nitrogen metabolism
13_ Amino acid metabolism
16_ Secondary metabolism
17_ Hormone metabolism
18_ Co-factor and vitamin metabolism
20_ Stress biotic
21_ Redox, thioredoxin
26_ Miscellaneous
27_ RNA
28_ DNA
29_ Protein
30_ Signalling
31_ Cell
33_ Development
34_ Transport
35_ Not assigned

Figure 4.5 A graphic representation of the major and minor co-expression gene networks consisted of 279 of the 303 genes (Appendix 2) that were associated with *Ma1*. The networks were constructed using the Cytoscape network inferring tool (Pearson's r>0.90) and analyzed with NetworkAnalyzer (Assenov et al. 2008). Node sizes are in proportion with their degree values. Node color and shape keys are indicated in the legend. **A.** A major co-expression gene network of 264 members. **B.** The minor networks. **C.** Distribution of node degrees in the major network.


A
B
Legend
Photosynthesis Bin 1
Transporter Bin 34
Transcription factor Bin 27.3
Receptor kinase, signaling Bin 29.4, 30
Respiration Bin 8, 9, 34.9
Hormone metabolism Bin 17
Biotic and Abiotic stress Bin 20
Up-regulated in MHA group
Down-regulated in MHA group
C
Number of nodes
Degree
0
5
10
15
20
25
30
35
40
0
4
8
12
16
20
24
28

which were listed in Table 4.3.

A close look at the major co-expression network revealed that *Ma1* had 13 primary neighbors (Figure 4.6A) and 78 secondary neighbors (Figure 4.6B). The expression of the 13 primary neighboring genes showed that ten of them were up-regulated and three were down-regulated in the *Ma__* genotype group in relation to the *mama* group (Table 4.3). The ten up-regulated genes included two (M140330 and M319170) putatively for encoding calcium signaling proteins, one (M682401) for pyruvate dehydrogenase in TCA cycle, one (M250124) for serine/threonine protein kinase, one (M911376) for phosphatidylcholine-sterol O-acyltransferase in lipid metabolism, and five for proteins of unknown function. The three down-regulated genes, however, putatively encoded a UDP-glucosyltransferase (M752561), a polyphenol oxidase (M744636) and a protein of unknown function (G200575) (Table 4.3).

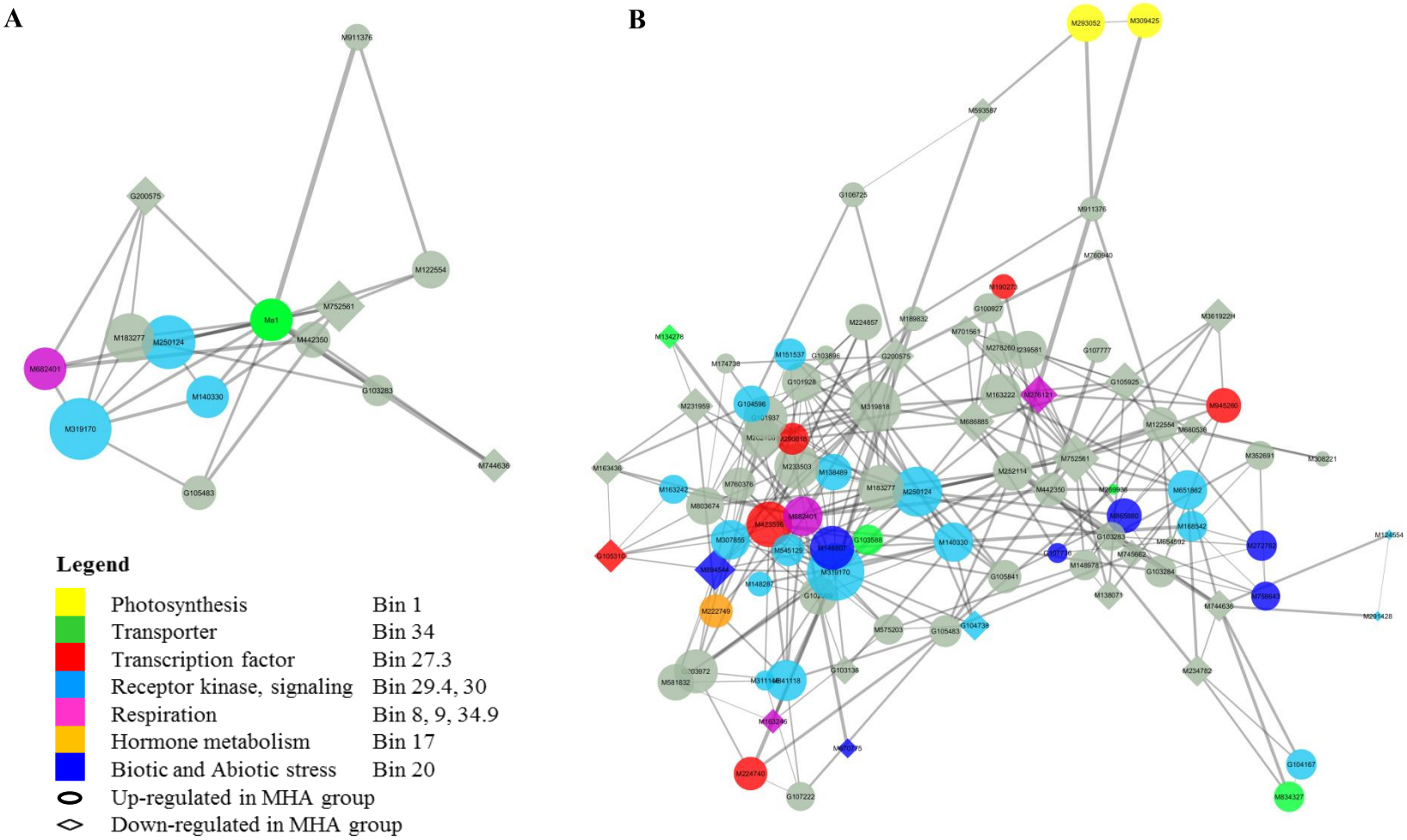
Among the 78 secondary neighboring genes of *Ma1*, 64 were functionally annotated by MapMan, including 44 up-regulated and 20 down-regulated in the *Ma__* group in relation to the *mama* group (Appendix 2). Notably, there were six transcription factors in the list, i.e. M945260 (an auxin responsive factor), M423596 (a homeobox factor), M224740 and M290818 (both similar to Arabidopsis response regulator1 (ARR1)), M190273 (an ethylene insensitive 3-like factor (EIL)), and G105310 (an ethylene responsive factor). The first five of the six transcription factors were up-regulated and the last one was down-regulated in the *Ma__* group. The secondary neighboring genes also included three transporter encoding genes that were up-regulated in the *Ma__* group, i.e. M834327 (a cyclic nucleotide gated channel), G103588 (an aquaporin PIP2) and M269936 (an ABC transporter). Given the limited scope of this study, the other secondary neighboring genes (Appendix 2) would not be described.

Table 4.3 List of genes of the highest node degrees in the *Ma1* associated co-expression network

Gene ID*	RPKM of *mama*	RPKM of *Ma__*	Network Degree	Order of *Ma1* neighbors	MapMan (sub)bin	Gene annotation
M319170	9.15	15.14	25	1st	30.3	IQ domain-containing protein, calmodulin binding
M319818	2.15	4.42	21	2nd	29.5.11.4.3.2	F-box protein, ubiquitin
M250124	4.68	9.63	20	1st	29.4	Serine/threonine protein kinase
M752561	4.66	0.32	19	1st	26.2	Indole-3-acetate beta-glucosyltransferase (EC 2.4.1.121), UDP-glycosyltransferase
M423596	6.53	15.80	17	2nd	27.3.22	Homeobox-leucine zipper protein, (ATHB13)
M233503	15.90	31.34	17	2nd	35.2	Putative uncharacterized protein RAF9-1
M183277	16.46	31.61	17	1st	35.2	Putative uncharacterized protein T21H19_30
M148807	10.84	16.06	16	2nd	20.2.3	Methyltransferase
G101937	8.86	14.35	16	2nd	31.4	SNARE-like superfamily protein, vesicle-mediated transport in plasma membrane
G203972	6.54	20.37	16	2nd	35.2	Unknown protein
M686885	950.26	532.89	14	2nd	13.1.5.2.41	Sarcosine oxidase, amino acid synthesis
M894544	110.04	81.61	14	2nd	20.2.99	Ozone-responsive stress related protein
M841118	1.35	3.27	14	2nd	30.3	Calcium-binding protein
M163222	3.72	10.14	14	2nd	31.1	Fimbrin, organisation
M273014	1.08	5.33	13	≥3rd	1.3.7	FBPase, Fructose-1,6-bisphosphatase (EC 3.1.3.11), calvin cycle
M682401	0.04	1.65	13	1st	8.1.1.2	Dihydrolipoyllysine-residue acetyltransferase in pyruvate dehydrogenase complex
M651862	2.21	5.76	13	2nd	29.4.1.57	BAK1, an LRR receptor kinase, binding BRI1, in brassinosteroid signaling
M307855	2.68	8.99	13	2nd	30.11	Phototropism protein, light signalling
M140330	2.94	6.78	13	1st	30.3	Calmodulin-like protein, calcium singalling
G105925	22.12	14.07	13	2nd	33.99	Protein with RING/U-box and TRAF-like domains, ubiquitin-protein ligase,
M252114	11.70	19.62	13	1st	34.8.1	Aluminum activated malate transporter
M202406	2.05	4.89	13	2nd	35.2	Desumoylating isopeptidase 2
G101928	18.67	30.65	13	2nd	35.2	Unknown protein
G200575	120.87	88.08	11	1st	35.2	Unknown protein
M122554	3.09	6.59	10	1st	35.2	Putative uncharacterized protein T2E22.10
M442350	0.60	5.46	9	1st	35.2	Unknown protein
M744636	2.95	0.62	8	1st	35.2	Polyphenol oxidase, chloroplastic
G105483	3.25	7.65	8	1st	35.2	Unknown protein
G103283	1.96	11.60	6	1st	35.2	Unknown protein
M911376	4.34	7.44	5	1st	11.8.10	Phosphatidylcholinesterol O-acyltransferase, lipid metabolism

* The letter ‘M’ in gene IDs is abbreviated from ‘MDP0000’ in the original apple gene IDs. Genes starting with ‘G’ are novel transcripts in the improved apple reference transcriptome (Bai et al. 2014)

Figure 4.6 A graphic representation of the *Ma1* associated co-expression gene networks consisted of 92 primary and secondary neighboring genes. The networks were constructed using the Cytoscape network inferring tool (Pearson's r>0.90) and analyzed with NetworkAnalyzer (Assenov et al. 2008). Node sizes were in proportion with their degree values. Node color and shape keys are indicated in the legend. **A.** A co-expression gene network of the primary neighbors of *Ma1* of 14 nodes. **B.** A co-expression gene network of the primary and secondary neighbors of *Ma1* of 92 nodes.


A
B
M911376
G200575
M122554
M752561
Ma1
M183277
M250124
M442350
M682401
M140330
G103283
M319170
M744636
G105483
Legend
Photosynthesis Bin 1
Transporter Bin 34
Transcription factor Bin 27.3
Receptor kinase, signaling Bin 29.4, 30
Respiration Bin 8, 9, 34.9
Hormone metabolism Bin 17
Biotic and Abiotic stress Bin 20
Up-regulated in MHA group
Down-regulated in MHA group

4.3.5 qRT-PCR confirmation of gene expression

To evaluate whether or not the RPKM values truly reflect the gene expression levels, eight genes, including *Ma1*, M190273, M651862, M815327, M132720, M163222, M196894 and M219042, were analyzed using qRT-PCR (Figure 4.7). The data confirmed that the relative expressions of the eight genes in qRT-PCR were significantly ($P<0.05$) correlated with their RPKM values in RNA-seq.

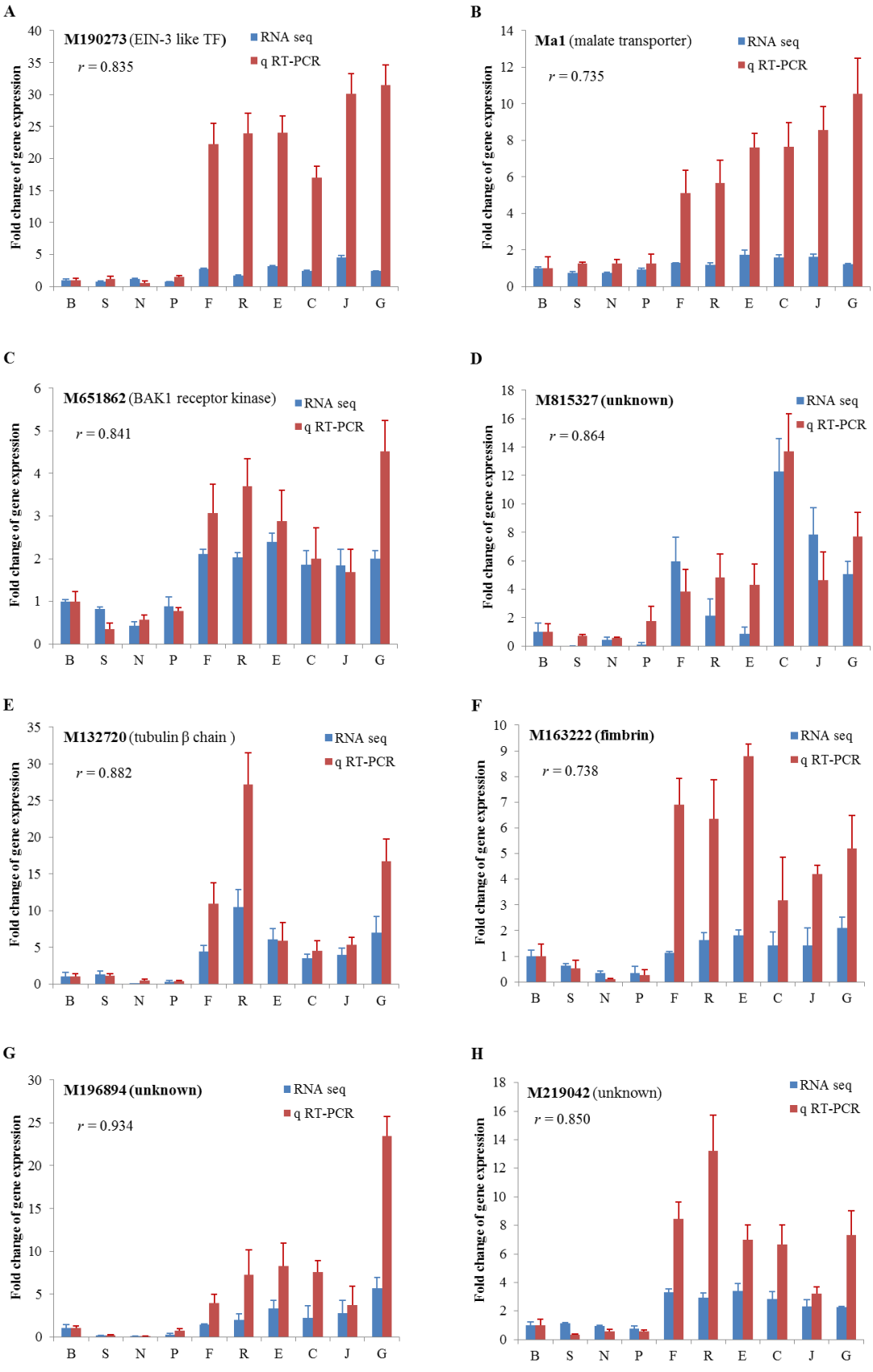
4.4 Discussion

4.4.1 Identification of genes expressed differentially between the two genotype groups *mama* and *Ma_*

In this study, ten apple varieties were carefully chosen to represent the two genotype groups *mama* and *Ma__* (*MaMa* and *Mama*) of contrast fruit acidity levels. These varieties were grafted on the same rootstock P22 and grown in the same orchard with the same management to curtail the environmental impacts on gene expression. However, since the apples are of different genetic background and are matured at varying dates, fruit transcriptome at maturity were inevitably influenced by these genetic and environmental variables. To minimize such bias, three replicates per variety and at least four varieties per genotype group were sampled and used for RNA-seq analysis. Although this strategy might not be perfect, we are convinced it is appropriate and adequate for the defined objectives in this study.

Overall, we found 38,348 genes expressed (RPKM >0.3) in the *mama* group and 39,854 genes in the *Ma__* group. Since there were 36,737 genes expressed in both

Figure 4.7 Confirmation of gene expression of eight selected genes using qRT-PCR. **A-H.** The normalized expression of target genes relative to a control gene (actin) in qRT-PCR was shown in red, and their corresponding RPKM values from RNA-seq were in blue. The correlation coefficient (*r*) was shown accordingly.


A
M190273 (EIN-3 like TF)
r = 0.835
RNA seq
q RT-PCR
Fold change of gene expression
B S N P F R E C J G
B
Ma1 (malate transporter)
r = 0.735
C
M651862 (BAK1 receptor kinase)
r = 0.841
D
M815327 (unknown)
r = 0.864
E
M132720 (tubulin β chain)
r = 0.882
F
M163222 (fimbrin)
r = 0.738
G
M196894 (unknown)
r = 0.934
H
M219042 (unknown)
r = 0.850

groups, the total number of expressed genes was 41,465. Of these expressed genes, 303 were not only expressed significantly (P_{FDR}<0.05) differentially between the *mama* and *Ma__* groups but also in significant ($P_{r>0.632}$<0.05) correlation with *Ma1* expression in the ten apple varieties. Considering the fact that the confirmation of significant correlation between the RNA seq and qRT-PCR data in the ten genes tested, we believe that the majority, if not all, of the genes that were responsible for higher expression of *Ma1* in genotype *Ma_* than in genotype *mama*, that were essential for the change of the transcriptome landscape between the two genotype groups and that were members important for the *Ma1* associated co-expression gene network governing fruit acidity were included in the 303 genes.

4.4.2 Regulation of the expression of *Ma1* and its associated co-expression gene network governing fruit acidity

The major co-expression gene network comprised 264 of the 303 genes associated with *Ma1*, of which 23 were of the highest degrees (13-25) (Figure 4.5). The degree threshold for defining genes of the 'highest degrees' in the network was equal to the degree 13 observed for *Ma1*. We considered this threshold was appropriate given the master regulatory role of *Ma1* in apple fruit acidity (Bai et al 2012). The gene that had the top degree (25) in the network was M319170, which encoded a calmodulin binding protein involving calcium signaling. Moreover, there were two additional genes in the gene list of the highest degrees (Table 4.3), i.e. M140330 (Calmodulin-like protein) and M841118 (encoding a calcium-dependent phospholipid binding protein) also involved in the calcium signaling. Further, these three calcium signaling related genes were all up-regulated in the *Ma_* genotype group in relation to the *mama* group, and M319170 and M140330 were two of the 13 primary neighbors of *Ma1*

while M841118 a secondary neighbor. These observations strongly suggested calcium signaling modulated by genes M319170, M140330 and M841118 was likely a crucial mechanism that regulates the expression of *Ma1* and its associated co-expression gene network governing fruit acidity.

The current model accounting for the ALMT1 (the first members of the ALMT gene family that includes *Ma1*) mediated plant tolerance to aluminum toxicity comprises a series of elements, including environmental cures (Al3+/H+), a receptor (unknown), signal transduction (unknown), kinase and/or phosphatase specific for transcriptional factor (unknown), transcription factors (STOP1 in Arabidopsis and ART1 in rice, both are C2H2 zinc fingers) and responsive genes (ALMT1 and others) (Delhaize et al. 2012). In a transcriptomics study in developing apple fruit (Bai et al. 2014a), several groups of genes of functionally similar to these elements were co-enriched or -suppressed in a malate dependable manner, including 14 C2H2 transcriptional factors, 27 protein kinase, and 23 receptor kinases for signaling. In addition, G105811, one of 14 C2H2 transcription factors, was even annotated as a STOP1-like protein by Mercator. These had led to a proposal for a transcriptional regulation model for malate variations in developing fruit similar to the model for aluminum tolerance, but there was no element potentially for the environmental cure speculated (Bai et al 2014b). In this study, a notion was clear that calcium signaling is likely vital for fruit acidity. Based on this notion, we propose that calcium be the intra-/inter-cellular cure and calmodulin (e.g. M140330 and M841118) be the receptor and for the fruit malate model. Indeed, in literature calcium (Ca^{2+}) has been regarded as one of the most prominent secondary messengers in plant (White and Broadley 2003) and calmodulins (CaM), a group of calcium sensor and signal transducer proteins, are known for their roles in regulating diverse cellular functions, such as gene expression, enzyme activity, and transport across membranes (Bouché et al. 2005). If this model

stood, M319170 the gene of the most degree in the network and encoding a calmodulin binding protein would be regulator for the calcium mediated signaling. It should be pointed out that, however, this study did not identify a single C2H2 zinc finger protein encoding gene that was associated with *Ma1*, an key element in the *ALMT1* mediated aluminum tolerance model although one transcription factor (M423596, a homeobox-leucine zipper protein) was found in the 23 genes of the highest degrees and four (M190273, M224740, M290818, M950387 and G105310) were listed in the secondary neighbors of *Ma1*.

4.4.3 Expression of other transporters in support of the vacuole 'acid trap' theory

Our previous study reported that *Ma1*, a putative aluminum activated malate transporter (ALMT) like gene, was the primary determinant of malate content in mature apple fruit (Bai et al. 2012). MA1 protein was highly homologous to *Arabidopsis* malate channel proteins AtALMT6 and AtALMT9 that were located on vacuolar membrane (Bai et al. 2012; Khan et al. 2012; Meyer et al. 2011). Berüter (2004) observed that uptake of ^{14}C-labelled malate was significantly lower in excised apple fruit of a low-acid genotype than in high-acid fruit and suggested that the reduced capacity of vacuolar storage in the low-acid fruit may lead to its higher rate of malate degradation. According to these findings, the vacuolar storage of malate may play an essential role in regulating fruit acidity in mature apple, while the malate related metabolism may respond to the vacuolar accumulation accordingly.

Based on the 'acid trap' theory in Martinoia et al. (2007) (2012), almost all malate was in the form of dianion in the cytosolic environment at neutral or slightly alkaline pH. Dianion malate was the only chemical form that malate transporters and channels can transport (Berüter 2004; Martinoia et al. 2007; Oleski et al. 1987). Once

the dianion malate crosses the tonoplast and enters the vacuole, where the pH turns to acidic, it is protonated and 'trapped' inside the vacuole, generating its electrochemical potential gradient ($\Delta\psi$). The trapping ability of malate depends on both the vacuolar pH and the malate $\Delta\psi$ across tonoplast. In this study, there were higher levels of malate accumulated in vacuole in the *Ma__* varieties than in the *mama* varieties. In the vacuoles of the apple cells in genotype *Ma__*, increased cation influx should be observed to neutralize the high concentration of anion (mostly malate). In the 303 genes, the expression of voltage-gated potassium channel (M322339), a cation translocator, was up-regulated in the *Ma__* varieties. Notably, the plasma aquaporin PIP2 (G103588) was expressed more than five times higher in the *Ma__* than in the *mama* apples, indicating that the cellular influx and efflux of H_2O might play a role in maintaining the vacuolar 'trapping' ability because aquaporins could adjust cytosolic malate concentration and thus the malate $\Delta\psi$ across tonoplast. In sum, not only *Ma1*, the aluminum activated malate transporter, but also the potassium channel and the aquaporin jointly contributed to the vacuolar accumulation of malate in apple fruit cells.

Summary By comparing the fruit transcriptomes from ten apple varieties in two genotype groups *mama* and *Ma_* (including *MaMa* and *Mama*) of contrast fruit acidity levels, a set of 303 genes was identified to be expressed not only differentially between the two groups but also in significant correlation with the expression of *Ma1*. Network inferring from the 303 genes revealed a major *Ma1* associated co-expression gene network of 264 nodes (genes). Analysis of the major network uncovered a subset of 23 genes of the high network degrees, which included *Ma1*. Within the major co-expression network, *Ma1* was characterized with 13 primary and 78 secondary neighboring genes. Based on the putative function of the 23 genes of the high network

degrees and/or the 11 primary neighboring gene of *Ma1*, we concluded that calcium signaling modulated by genes M319170 (encoding a calmodulin binding protein), M140330 (encoding a calmodulin-like protein) and M841118 (encoding a calcium-dependent phospholipid binding protein) was likely a crucial mechanism that regulates both the expression of *Ma1* and the *Ma1* associated co-expression gene network governing fruit acidity.

4.5 Reference

Assenov Y, Ramírez F, Schelhorn S-E, Lengauer T, Albrecht M (2008) Computing topological parameters of biological networks. Bioinformatics 24:282-284

Bai Y, Dougherty L, cheng L, Xu K (2014a) A co-expression gene network regulating acidity in developing apple fruit

Bai Y, Dougherty L, Li M, Fazio G, Cheng L, Xu K (2012) A natural mutation-led truncation in one of the two aluminum-activated malate transporter-like genes at the Ma locus is associated with low fruit acidity in apple. Molecular Genetics and Genomics 287: 663-678

Bai Y, Dougherty L, Xu K (2014b) Towards an improved apple reference transcriptome using RNA-seq. Mol Genet Genomics

Benjamini Y, Hochberg Y (1995) Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. J R Stat Soc Ser B-Methodol 57:289-300

Berüter J (2004) Carbohydrate metabolism in two apple genotypes that differ in malate accumulation. Journal of Plant Physiology 161:1011-1029

Blanpied GD, Silsby KJ (1992) Predicting harvest date windows for apples. Information Bulletin 221. Cornell Cooperative Extension, Cornell University, Ithaca, NY, USA.

Bouché N, Yellin A, Snedden WA, Fromm H (2005) PLANT-SPECIFIC CALMODULIN-BINDING PROTEINS. Annual Review of Plant Biology 56:435-466

Delhaize E, Ma JF, Ryan PR (2012) Transcriptional regulation of aluminium tolerance genes. Trends in Plant Science doi: 10.1016/j.tplants.2012.02.008

Hulme AC, Wooltorton LSC (1957) The organic acid metabolism of apple fruits: changes in individual acids during growth on the tree. Journal of the Science of Food and Agriculture 8:117-122

Kang C, Darwish O, Geretz A, Shahan R, Alkharouf N, Liu Z (2013) Genome-scale transcriptomic insights into early-stage fruit development in woodland strawberry Fragaria vesca. Plant Cell 25:1960-1978

Kenis K, Keulemans J, Davey M (2008) Identification and stability of QTLs for fruit quality traits in apple. Tree Genetics & Genomes 4:647-661

Khan SA, Beekwilder J, Schaart JG, Mumm R, Soriano JM, Jacobsen E, Schouten HJ (2012) Differences in acidity of apples are probably mainly caused by a malic acid transporter gene on LG16. Tree Genetics & Genomes 9:475-487

Krost C, Petersen R, Lokan S, Brauksiepe B, Braun P, Schmidt E (2013) Evaluation of the hormonal state of columnar apple trees (Malus x domestica) based on high throughput gene expression studies. Plant Molecular Biology 81:211-220

Krost C, Petersen R, Schmidt ER (2012) The transcriptomes of columnar and standard type apple trees (Malus x domestica) - A comparative study. Gene 498:223-230

Kumar S, Chagne D, Bink MCAM, Volz RK, Whitworth C, Carlisle C (2012) Genomic Selection for Fruit Quality Traits in Apple (Malus x domestica Borkh.). PLoS One 7:e36674

Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10:R25

Li P, Ponnala L, Gandotra N, Wang L, Si Y, Tausta SL, Kebrom TH, Provart N, Patel R, Myers CR, Reidel EJ, Turgeon R, Liu P, Sun Q, Nelson T, Brutnell TP (2010) The developmental dynamics of the maize leaf transcriptome. Nat Genet 42:1060-1067

Liebhard R, Kellerhals M, Pfammatter W, Jertmini M, Gessler C (2003) Mapping quantitative physiological traits in apple (Malus x domestica Borkh.). Plant Molecular Biology 52:511-526

Lister R, O'Malley RC, Tonti-Filippini J, Gregory BD, Berry CC, Millar AH, Ecker JR (2008) Highly integrated single-base resolution maps of the epigenome in Arabidopsis. Cell 133:523-536

M R-C, S Y, M Z, H F, W G (1999) The prenylation status of a novel plant calmodulin directs plasma membrane or nuclear localization of the protein. EMBO J 18:1996

Maliepaard C, Alston FH, van Arkel G, Brown LM, Chevreau E, Dunemann F, Evans KM, Gardiner S, Guilford P, van Heusden AW, Janse J, Laurens F, Lynn JR, Manganaris AG, den Nijs APM, Periam N, Rikkerink E, Roche P, Ryder C, Sansavini S, Schmidt H, Tartarini S, Verhaegh JJ, Vrielink-van Ginkel M, King GJ (1998) Aligning male and female linkage maps of apple (Malus pumila Mill.) using multi-allelic markers. Theoretical and Applied Genetics 97:60-73

Martinoia E, Maeshima M, Neuhaus HE (2007) Vacuolar transporters and their essential role in plant metabolism. J Exp Bot 58:83-102

Martinoia E, Meyer S, De Angeli A, Nagy Rk (2012) Vacuolar Transporters in Their Physiological Context. Annual Review of Plant Biology 63:183-213

Meyer S, Scholz-Starke J, De Angeli A, Kovermann P, Burla B, Gambale F, Martinoia E (2011) Malate transport by the vacuolar AtALMT6 channel in guard cells is subject to multiple regulation. Plant Journal 67:247-257

MJ C, PL V, R K, SH L, JP D (1998) Reciprocal regulation of mammalian nitric oxide synthase and calcineurin by plant calmodulin isoforms. Biochemistry 37:15593

Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nature Methods 5:621-628

Oleski N, Mahdavi P, Bennett AB (1987) Transport Properties of the Tomato Fruit Tonoplast: II. Citrate Transport. Plant Physiology 84:997-1000

Pfaffl MW (2001) A new mathematical model for relative quantification in real-time RT-PCR. Nucleic Acids Research 29:e45

Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glockner FO (2013) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic acids research 41:D590-596

Saito R, Smoot ME, Ono K, Ruscheinski J, Wang P-L, Lotia S, Pico AR, Bader GD, Ideker T (2012) A travel guide to Cytoscape plugins. Nat Meth 9:1069-1076

T A, G B, WA S, BJ S, H F (1995) Molecular and biochemical analysis of calmodulin interactions with the calmodulin-binding domain of plant glutamate decarboxylase. Plant Physiol 108:551

White PJ, Broadley MR (2003) Calcium in Plants. Annals of Botany 92:487-511

Wilhelm BT, Landry JR (2009) RNA-Seq-quantitative measurement of expression through massively parallel RNA-sequencing. Methods 48:249-257

Xia R, Zhu H, An Y-q, Beers E, Liu Z (2012) Apple miRNAs and tasiRNAs with novel regulatory networks. Genome Biology 13:R47

Xu K, Wang A, Brown S (2012) Genetic characterization of the Ma locus with pH and titratable acidity in apple. Molecular Breeding 30:899-912

Yamaki S (1984) Isolation of vacuoles from immature apple fruit flesh and compartmentation of sugars, organic acids, phenolic compounds and amino acids. Plant Cell Physiol 25:151-166

Zenoni S, Ferrarini A, Giacomelli E, Xumerle L, Fasoli M, Malerba G, Bellin D, Pezzotti M, Delledonne M (2010) Characterization of Transcriptional Complexity during Berry Development in Vitis vinifera Using RNA-Seq. Plant Physiology 152:1787-1795

Zhang Q, Ma B, Li H, Chang Y, Han Y, Li J, Wei G, Zhao S, Khan M, Zhou Y, Gu C, Zhang X, Han Z, Korban S, Li S, Han Y (2012a) Identification, characterization, and

utilization of genome-wide simple sequence repeats to identify a QTL for acidity in apple. BMC Genomics 13:537

Zhang Y, Zhu J, Dai H (2012b) Characterization of Transcriptional Differences Between Columnar and Standard Apple Trees Using RNA-Seq. Plant Molecular Biology Reporter 30:957-965

Zhang YZ, Li PM, Cheng LL (2010) Developmental changes of carbohydrates, organic acids, amino acids, and phenolic compounds in 'Honeycrisp' apple flesh. Food Chemistry 123:1013-1018

Appendix 1 List of genes in the co-expression network regulating malate levels in developing fruit of Golden Delicious

Gene ID*	K-Means Cluster	Network Degree	Order of neighbors	MapMan (sub)bin	Bin description
M177322	2	10	2nd	1.1.1.2	PS.lightreaction.photosystem II.PSII polypeptide subunits
M647505	2	14	2nd	1.1.1.2	PS.lightreaction.photosystem II.PSII polypeptide subunits
M764147	3	16	2nd	1.1.1.2	PS.lightreaction.photosystem II.PSII polypeptide subunits
M274141	1	8	≥3rd	1.1.1.2	PS.lightreaction.photosystem II.PSII polypeptide subunits
M906067	1	10	≥3rd	1.1.1.2	PS.lightreaction.photosystem II.PSII polypeptide subunits
M132424	1	7	≥3rd	1.1.1.2	PS.lightreaction.photosystem II.PSII polypeptide subunits
M132436	1	6	≥3rd	1.1.1.2	PS.lightreaction.photosystem II.PSII polypeptide subunits
M258622	1	13	≥3rd	1.1.1.2	PS.lightreaction.photosystem II.PSII polypeptide subunits
M217362	1	10	≥3rd	1.1.1.2	PS.lightreaction.photosystem II.PSII polypeptide subunits
M248920	1	11	≥3rd	1.1.1.2	PS.lightreaction.photosystem II.PSII polypeptide subunits
M805790	1	5	≥3rd	1.1.1.2	PS.lightreaction.photosystem II.PSII polypeptide subunits
G303514	1	2	≥3rd	1.1.1.2	PS.lightreaction.photosystem II.PSII polypeptide subunits
M538720	1	4	≥3rd	1.1.1.2	PS.lightreaction.photosystem II.PSII polypeptide subunits
M858039	1	13	≥3rd	1.1.1.2	PS.lightreaction.photosystem II.PSII polypeptide subunits
M281971	1	14	≥3rd	1.1.1.2	PS.lightreaction.photosystem II.PSII polypeptide subunits
M208621	1	12	≥3rd	1.1.1.2	PS.lightreaction.photosystem II.PSII polypeptide subunits
M195878	1	14	≥3rd	1.1.1.2	PS.lightreaction.photosystem II.PSII polypeptide subunits
M127699	1	8	≥3rd	1.1.1.2	PS.lightreaction.photosystem II.PSII polypeptide subunits
M203627	1	6	≥3rd	1.1.1.2	PS.lightreaction.photosystem II.PSII polypeptide subunits
M260399	1	1	≥3rd	1.1.1.2	PS.lightreaction.photosystem II.PSII polypeptide subunits
M451103	3	6	≥3rd	1.1.1.2	PS.lightreaction.photosystem II.PSII polypeptide subunits
M125631	1	1	NA	1.1.1.2	PS.lightreaction.photosystem II.PSII polypeptide subunits
M481445	1	1	NA	1.1.1.2	PS.lightreaction.photosystem II.PSII polypeptide subunits
M430367	3	1	NA	1.1.1.2	PS.lightreaction.photosystem II.PSII polypeptide subunits
M596458	3	1	NA	1.1.1.2	PS.lightreaction.photosystem II.PSII polypeptide subunits
M443024	1	NA	NA	1.1.1.2	PS.lightreaction.photosystem II.PSII polypeptide subunits
M154436	1	NA	NA	1.1.1.2	PS.lightreaction.photosystem II.PSII polypeptide subunits
M785123	1	NA	NA	1.1.1.2	PS.lightreaction.photosystem II.PSII polypeptide subunits

Appendix 1 Continued

Gene ID	K-Means Cluster	Network Degree	Order of neighbors	MapMan (sub)bin	Bin description
M132802	3	NA	NA	1.1.1.2	PS.lightreaction.photosystem II.PSII polypeptide subunits
M624964	3	15	≥3rd	1.1.2.2	PS.lightreaction.photosystem I.PSI polypeptide subunits
G200008	1	6	≥3rd	1.1.2.2	PS.lightreaction.photosystem I.PSI polypeptide subunits
M128789	1	1	≥3rd	1.1.2.2	PS.lightreaction.photosystem I.PSI polypeptide subunits
M154360	1	12	≥3rd	1.1.2.2	PS.lightreaction.photosystem I.PSI polypeptide subunits
M452016	1	9	≥3rd	1.1.2.2	PS.lightreaction.photosystem I.PSI polypeptide subunits
M700880	1	4	≥3rd	1.1.2.2	PS.lightreaction.photosystem I.PSI polypeptide subunits
M128787	1	7	≥3rd	1.1.2.2	PS.lightreaction.photosystem I.PSI polypeptide subunits
M150637	1	8	≥3rd	1.1.2.2	PS.lightreaction.photosystem I.PSI polypeptide subunits
M909197	1	14	≥3rd	1.1.2.2	PS.lightreaction.photosystem I.PSI polypeptide subunits
M277240	1	10	≥3rd	1.1.2.2	PS.lightreaction.photosystem I.PSI polypeptide subunits
M184046	1	9	≥3rd	1.1.2.2	PS.lightreaction.photosystem I.PSI polypeptide subunits
M319772	1	8	≥3rd	1.1.2.2	PS.lightreaction.photosystem I.PSI polypeptide subunits
M142895	1	2	≥3rd	1.1.2.2	PS.lightreaction.photosystem I.PSI polypeptide subunits
M122042	1	8	≥3rd	1.1.2.2	PS.lightreaction.photosystem I.PSI polypeptide subunits
M237036	1	8	≥3rd	1.1.2.2	PS.lightreaction.photosystem I.PSI polypeptide subunits
M773525	1	6	≥3rd	1.1.2.2	PS.lightreaction.photosystem I.PSI polypeptide subunits
M481097	1	7	≥3rd	1.1.2.2	PS.lightreaction.photosystem I.PSI polypeptide subunits
M210616	1	2	≥3rd	1.1.2.2	PS.lightreaction.photosystem I.PSI polypeptide subunits
M268037	2	35	2nd	8.2.10	TCA / org transformation.other organic acid transformatons.malic
M221561	2	16	2nd	8.2.10	TCA / org transformation.other organic acid transformatons.malic
M376988	2	30	2nd	8.2.10	TCA / org transformation.other organic acid transformatons.malic
G200077	2	19	2nd	8.2.10	TCA / org transformation.other organic acid transformatons.malic
G200076	2	11	2nd	8.2.10	TCA / org transformation.other organic acid transformatons.malic
M258977	2	44	1st	8.2.10	TCA / org transformation.other organic acid transformatons.malic
G200225	2	5	≥3rd	8.2.10	TCA / org transformation.other organic acid transformatons.malic
M291902	2	14	≥3rd	8.2.10	TCA / org transformation.other organic acid transformatons.malic
M132833	2	NA	NA	8.2.10	TCA / org transformation.other organic acid transformatons.malic

Appendix 1 Continued

Gene ID	K-Means Cluster	Network Degree	Order of neighbors	MapMan (sub)bin	Bin description
M256239	2	28	2nd	9.1.2	mitochondrial electron transport / ATP synthesis.NADH-DH.localisation not clear
M278019	2	12	2nd	9.1.2	mitochondrial electron transport / ATP synthesis.NADH-DH.localisation not clear
M313179	2	NA	NA	9.1.2	mitochondrial electron transport / ATP synthesis.NADH-DH.localisation not clear
M153587	2	11	≥3rd	9.5	mitochondrial electron transport / ATP synthesis.cytochrome c reductase
M763693	3	NA	NA	9.5	mitochondrial electron transport / ATP synthesis.cytochrome c reductase
M945182	2	19	2nd	9.9	mitochondrial electron transport / ATP synthesis.F1-ATPase
M655872	2	2	≥3rd	9.9	mitochondrial electron transport / ATP synthesis.F1-ATPase
M360515	2	2	≥3rd	9.9	mitochondrial electron transport / ATP synthesis.F1-ATPase
M390327	2	NA	NA	9.9	mitochondrial electron transport / ATP synthesis.F1-ATPase
M131397	4	30	2nd	10.6.1	cell wall.degradation.cellulases and beta -1,4-glucanases
M252536	4	42	1st	10.6.1	cell wall.degradation.cellulases and beta -1,4-glucanases
M147635	5	2	≥3rd	10.6.1	cell wall.degradation.cellulases and beta -1,4-glucanases
M276676	6	9	≥3rd	10.6.1	cell wall.degradation.cellulases and beta -1,4-glucanases
M321839	4	40	1st	10.6.2	cell wall.degradation.mannan-xylose-arabinose-fucose
M319156	4	26	2nd	10.6.3	cell wall.degradation.pectate lyases and polygalacturonases
M631698	4	15	2nd	10.6.3	cell wall.degradation.pectate lyases and polygalacturonases
M236210	4	28	2nd	10.6.3	cell wall.degradation.pectate lyases and polygalacturonases
M491898	4	48	1st	10.6.3	cell wall.degradation.pectate lyases and polygalacturonases
M943790	3	8	≥3rd	10.6.3	cell wall.degradation.pectate lyases and polygalacturonases
M251956	3	2	≥3rd	10.6.3	cell wall.degradation.pectate lyases and polygalacturonases
M232225	5	3	≥3rd	10.6.3	cell wall.degradation.pectate lyases and polygalacturonases
M394944	6	1	≥3rd	10.6.3	cell wall.degradation.pectate lyases and polygalacturonases
M181608	6	4	≥3rd	10.6.3	cell wall.degradation.pectate lyases and polygalacturonases
M228673	2	NA	NA	10.6.3	cell wall.degradation.pectate lyases and polygalacturonases
M277149	5	NA	NA	10.6.3	cell wall.degradation.pectate lyases and polygalacturonases
M573746	5	NA	NA	10.6.3	cell wall.degradation.pectate lyases and polygalacturonases
M581832	5	NA	NA	10.6.3	cell wall.degradation.pectate lyases and polygalacturonases
G101335	1	NA	NA	13.2.3.1.1	amino acid metabolism.degradation.aspartate family.asparagine.L-asparaginase

Appendix 1 Continued

Gene ID	K-Means Cluster	Network Degree	Order of neighbors	MapMan (sub)bin	Bin description
M203071	2	NA	NA	13.2.3.2	amino acid metabolism.degradation.aspartate family.threonine
M212365	2	9	≥3rd	13.2.3.4	amino acid metabolism.degradation.aspartate family.methionine
M130701	2	12	≥3rd	13.2.3.5	amino acid metabolism.degradation.aspartate family.lysine
M129664	2	30	2nd	13.2.4.1	amino acid metabolism.degradation.branched chain group.shared
G200983	2	24	≥3rd	13.2.4.1	amino acid metabolism.degradation.branched chain group.shared
M161243	2	16	2nd	13.2.6.2	amino acid metabolism.degradation.aromatic aa.tyrosine
M934476	2	4	2nd	13.2.6.2	amino acid metabolism.degradation.aromatic aa.tyrosine
M180890	2	10	2nd	13.2.6.2	amino acid metabolism.degradation.aromatic aa.tyrosine
M134935	2	24	2nd	13.2.6.2	amino acid metabolism.degradation.aromatic aa.tyrosine
M258402	2	2	≥3rd	13.2.6.3	amino acid metabolism.degradation.aromatic aa.tryptophan
M237124	2	NA	NA	16.1.4.1	secondary metabolism.isoprenoids.carotenoids.phytoene synthase
G201485	2	13	2nd	16.1.4.10	secondary metabolism.isoprenoids.carotenoids.carotenoid cleavage dioxygenase
G105440	2	1	≥3rd	16.1.4.10	secondary metabolism.isoprenoids.carotenoids.carotenoid cleavage dioxygenase
M241703	2	5	≥3rd	16.1.4.2	secondary metabolism.isoprenoids.carotenoids.phytoene dehydrogenase
M255025	2	5	2nd	16.1.4.3	secondary metabolism.isoprenoids.carotenoids.zeta-carotene desaturase
M308095	2	7	2nd	16.1.4.3	secondary metabolism.isoprenoids.carotenoids.zeta-carotene desaturase
M145663	2	3	≥3rd	16.1.4.5	secondary metabolism.isoprenoids.carotenoids.lycopene beta cyclase
M225539	2	3	≥3rd	16.1.4.6	secondary metabolism.isoprenoids.carotenoids.carotenoid beta ring hydroxylase
G203831	4	23	2nd	16.8.1.1	secondary metabolism.flavonoids.anthocyanins.leucocyanidin dioxygenase
M754878	4	19	2nd	16.8.1.2	secondary metabolism.flavonoids.anthocyanins.anthocyanidin reductase
M126567	4	23	2nd	16.8.2	secondary metabolism.flavonoids.chalcones
M343259	4	25	2nd	16.8.2	secondary metabolism.flavonoids.chalcones
G202130	4	26	2nd	16.8.2	secondary metabolism.flavonoids.chalcones
M252589	4	33	2nd	16.8.2	secondary metabolism.flavonoids.chalcones
M134791	4	32	2nd	16.8.2	secondary metabolism.flavonoids.chalcones
M686666	4	20	2nd	16.8.2	secondary metabolism.flavonoids.chalcones
M575740	4	31	2nd	16.8.2	secondary metabolism.flavonoids.chalcones
G201827	4	20	2nd	16.8.2	secondary metabolism.flavonoids.chalcones

Appendix 1 Continued

Gene ID	K-Means Cluster	Network Degree	Order of neighbors	MapMan (sub)bin	Bin description
M361563	4	15	2nd	16.8.2	secondary metabolism.flavonoids.chalcones
M686661	4	24	2nd	16.8.2	secondary metabolism.flavonoids.chalcones
M523487	2	2	≥3rd	16.8.2	secondary metabolism.flavonoids.chalcones
M215073	2	NA	NA	16.8.2	secondary metabolism.flavonoids.chalcones
M494976	4	27	2nd	16.8.3.1	secondary metabolism.flavonoids.dihydroflavonols.dihydroflavonol 4-reductase
M243194	4	22	2nd	16.8.3.1	secondary metabolism.flavonoids.dihydroflavonols.dihydroflavonol 4-reductase
M243196	4	28	2nd	16.8.3.1	secondary metabolism.flavonoids.dihydroflavonols.dihydroflavonol 4-reductase
M271553	4	22	2nd	16.8.3.1	secondary metabolism.flavonoids.dihydroflavonols.dihydroflavonol 4-reductase
M320264	4	17	2nd	16.8.3.1	secondary metabolism.flavonoids.dihydroflavonols.dihydroflavonol 4-reductase
M174748	4	20	2nd	16.8.3.1	secondary metabolism.flavonoids.dihydroflavonols.dihydroflavonol 4-reductase
M607969	2	35	2nd	16.8.3.1	secondary metabolism.flavonoids.dihydroflavonols.dihydroflavonol 4-reductase
M268045	2	15	2nd	16.8.3.1	secondary metabolism.flavonoids.dihydroflavonols.dihydroflavonol 4-reductase
M322755	2	2	≥3rd	16.8.3.1	secondary metabolism.flavonoids.dihydroflavonols.dihydroflavonol 4-reductase
M706020	3	20	≥3rd	16.8.3.1	secondary metabolism.flavonoids.dihydroflavonols.dihydroflavonol 4-reductase
M320534	3	4	≥3rd	16.8.3.1	secondary metabolism.flavonoids.dihydroflavonols.dihydroflavonol 4-reductase
M307124	5	NA	NA	16.8.3.1	secondary metabolism.flavonoids.dihydroflavonols.dihydroflavonol 4-reductase
M202437	5	NA	NA	16.8.3.1	secondary metabolism.flavonoids.dihydroflavonols.dihydroflavonol 4-reductase
M220986	6	NA	NA	16.8.3.1	secondary metabolism.flavonoids.dihydroflavonols.dihydroflavonol 4-reductase
G104117	4	35	2nd	16.8.3.2	secondary metabolism.flavonoids.dihydroflavonols.flavanone 3-hydroxylase
G104118	4	12	≥3rd	16.8.3.2	secondary metabolism.flavonoids.dihydroflavonols.flavanone 3-hydroxylase
M381510	3	10	2nd	20.1.7	stress.biotic.PR-proteins
M168665	5	5	2nd	20.1.7	stress.biotic.PR-proteins
M245324	2	1	2nd	20.1.7	stress.biotic.PR-proteins
M650624	4	32	2nd	20.1.7	stress.biotic.PR-proteins
M139844	4	13	2nd	20.1.7	stress.biotic.PR-proteins
M520051	4	26	2nd	20.1.7	stress.biotic.PR-proteins
M293686	4	23	2nd	20.1.7	stress.biotic.PR-proteins
M261120	1	2	≥3rd	20.1.7	stress.biotic.PR-proteins

Appendix 1 Continued

Gene ID	K-Means Cluster	Network Degree	Order of neighbors	MapMan (sub)bin	Bin description
M252012	6	3	≥3rd	20.1.7	stress.biotic.PR-proteins
G100107	2	1	NA	20.1.7	stress.biotic.PR-proteins
M830772	3	NA	NA	20.1.7	stress.biotic.PR-proteins
M742278	4	18	2nd	20.2.99	stress.abiotic.unspecified
M570235	2	18	2nd	20.2.99	stress.abiotic.unspecified
M286324	2	11	2nd	20.2.99	stress.abiotic.unspecified
M310335	2	23	2nd	20.2.99	stress.abiotic.unspecified
M770493	2	34	2nd	20.2.99	stress.abiotic.unspecified
M427722	3	18	2nd	20.2.99	stress.abiotic.unspecified
M292853	3	21	2nd	20.2.99	stress.abiotic.unspecified
M216647	3	25	2nd	20.2.99	stress.abiotic.unspecified
M480605	3	21	2nd	20.2.99	stress.abiotic.unspecified
M870405	3	30	2nd	20.2.99	stress.abiotic.unspecified
M458276	4	34	2nd	20.2.99	stress.abiotic.unspecified
M652768	4	34	2nd	20.2.99	stress.abiotic.unspecified
M219522	4	37	2nd	20.2.99	stress.abiotic.unspecified
M513191	4	25	2nd	20.2.99	stress.abiotic.unspecified
M295908	4	42	1st	20.2.99	stress.abiotic.unspecified
G100001	4	41	1st	20.2.99	stress.abiotic.unspecified
M266399	3	3	≥3rd	20.2.99	stress.abiotic.unspecified
M151829	3	7	≥3rd	20.2.99	stress.abiotic.unspecified
M942514	3	10	≥3rd	20.2.99	stress.abiotic.unspecified
M520923	3	1	≥3rd	20.2.99	stress.abiotic.unspecified
M218018	5	2	≥3rd	20.2.99	stress.abiotic.unspecified
M707712	6	3	≥3rd	20.2.99	stress.abiotic.unspecified
M205674	6	14	≥3rd	20.2.99	stress.abiotic.unspecified
M530457	6	4	≥3rd	20.2.99	stress.abiotic.unspecified
M302671	6	7	≥3rd	20.2.99	stress.abiotic.unspecified

Appendix 1 Continued

Gene ID	K-Means Cluster	Network Degree	Order of neighbors	MapMan (sub)bin	Bin description
M397306	3	NA	NA	20.2.99	stress.abiotic.unspecified
M409061	3	NA	NA	20.2.99	stress.abiotic.unspecified
M725994	6	NA	NA	20.2.99	stress.abiotic.unspecified
M215774	4	12	2nd	26.21	misc.protease inhibitor/seed storage/lipid transfer protein (LTP) family protein
M835003	4	42	1st	26.21	misc.protease inhibitor/seed storage/lipid transfer protein (LTP) family protein
M293049	2	2	≥3rd	26.21	misc.protease inhibitor/seed storage/lipid transfer protein (LTP) family protein
M265942	3	2	≥3rd	26.21	misc.protease inhibitor/seed storage/lipid transfer protein (LTP) family protein
G302562	3	2	≥3rd	26.21	misc.protease inhibitor/seed storage/lipid transfer protein (LTP) family protein
M155446	5	2	≥3rd	26.21	misc.protease inhibitor/seed storage/lipid transfer protein (LTP) family protein
M781178	6	9	≥3rd	26.21	misc.protease inhibitor/seed storage/lipid transfer protein (LTP) family protein
M125336	1	NA	NA	26.21	misc.protease inhibitor/seed storage/lipid transfer protein (LTP) family protein
M324249	3	NA	NA	26.21	misc.protease inhibitor/seed storage/lipid transfer protein (LTP) family protein
M921319	5	NA	NA	26.21	misc.protease inhibitor/seed storage/lipid transfer protein (LTP) family protein
G105715	2	3	2nd	27.3.11	RNA.regulation of transcription.C2H2 zinc finger family
M171430	6	8	2nd	27.3.11	RNA.regulation of transcription.C2H2 zinc finger family
M183099	3	10	2nd	27.3.11	RNA.regulation of transcription.C2H2 zinc finger family
M441454	4	31	2nd	27.3.11	RNA.regulation of transcription.C2H2 zinc finger family
M427068	6	22	2nd	27.3.11	RNA.regulation of transcription.C2H2 zinc finger family
G104390	4	30	2nd	27.3.11	RNA.regulation of transcription.C2H2 zinc finger family
M246502	3	50	1st	27.3.11	RNA.regulation of transcription.C2H2 zinc finger family
M804707	2	2	≥3rd	27.3.11	RNA.regulation of transcription.C2H2 zinc finger family
G105811	5	3	≥3rd	27.3.11	RNA.regulation of transcription.C2H2 zinc finger family
M144266	6	6	≥3rd	27.3.11	RNA.regulation of transcription.C2H2 zinc finger family
M217209	1	NA	NA	27.3.11	RNA.regulation of transcription.C2H2 zinc finger family
M315201	2	NA	NA	27.3.11	RNA.regulation of transcription.C2H2 zinc finger family
G104902	3	NA	NA	27.3.11	RNA.regulation of transcription.C2H2 zinc finger family
M154566	6	NA	NA	27.3.11	RNA.regulation of transcription.C2H2 zinc finger family
M597906	1	7	≥3rd	29.2.1.1.1.1.17	protein.synthesis.ribosomal protein.prokaryotic.chloroplast.30S subunit.S17

Appendix 1 Continued

Gene ID	K-Means Cluster	Network Degree	Order of neighbors	MapMan (sub)bin	Bin description
G200032	2	NA	NA	29.2.1.1.1.1.3	protein.synthesis.ribosomal protein.prokaryotic.chloroplast.30S subunit.S3
M796276	1	11	≥3rd	29.2.1.1.1.1.31	protein.synthesis.ribosomal protein.prokaryotic.chloroplast.30S subunit.S31
M164974	1	3	≥3rd	29.2.1.1.1.1.9	protein.synthesis.ribosomal protein.prokaryotic.chloroplast.30S subunit.S9
M238515	3	3	≥3rd	29.2.1.1.1.2.11	protein.synthesis.ribosomal protein.prokaryotic.chloroplast.50S subunit.L11
M767733	2	NA	NA	29.2.1.1.1.2.12	protein.synthesis.ribosomal protein.prokaryotic.chloroplast.50S subunit.L12
M194017	3	4	≥3rd	29.2.1.1.1.2.14	protein.synthesis.ribosomal protein.prokaryotic.chloroplast.50S subunit.L14
M361249	2	NA	NA	29.2.1.1.1.2.14	protein.synthesis.ribosomal protein.prokaryotic.chloroplast.50S subunit.L14
M736369	3	1	≥3rd	29.2.1.1.1.2.15	protein.synthesis.ribosomal protein.prokaryotic.chloroplast.50S subunit.L15
M361241	2	NA	NA	29.2.1.1.1.2.16	protein.synthesis.ribosomal protein.prokaryotic.chloroplast.50S subunit.L16
M279154	3	15	2nd	29.2.1.1.1.2.18	protein.synthesis.ribosomal protein.prokaryotic.chloroplast.50S subunit.L18
M834632	3	11	≥3rd	29.2.1.1.1.2.18	protein.synthesis.ribosomal protein.prokaryotic.chloroplast.50S subunit.L18
M436957	2	NA	NA	29.2.1.1.1.2.2	protein.synthesis.ribosomal protein.prokaryotic.chloroplast.50S subunit.L2
M855100	3	5	≥3rd	29.2.1.1.1.2.34	protein.synthesis.ribosomal protein.prokaryotic.chloroplast.50S subunit.L34
M215992	3	NA	NA	29.2.1.1.1.2.34	protein.synthesis.ribosomal protein.prokaryotic.chloroplast.50S subunit.L34
M299779.1	2	11	≥3rd	29.2.1.1.1.2.4	protein.synthesis.ribosomal protein.prokaryotic.chloroplast.50S subunit.L4
M299779.2	2	10	≥3rd	29.2.1.1.1.2.4	protein.synthesis.ribosomal protein.prokaryotic.chloroplast.50S subunit.L4
M260620	3	2	≥3rd	29.2.1.1.1.2.4	protein.synthesis.ribosomal protein.prokaryotic.chloroplast.50S subunit.L4
M890106	3	1	≥3rd	29.2.1.1.1.2.86	protein.synthesis.ribosomal protein.prokaryotic.chloroplast.50S subunit.PSRP6
M303332	1	4	≥3rd	29.3.1	protein.targeting.nucleus
M177132	2	12	2nd	29.3.2	protein.targeting.mitochondria
M477455	2	4	≥3rd	29.3.2	protein.targeting.mitochondria
M309673	2	7	≥3rd	29.3.2	protein.targeting.mitochondria
M830613	2	NA	NA	29.3.2	protein.targeting.mitochondria
M171639	2	32	2nd	29.3	protein.targeting
M155675	2	10	≥3rd	29.3.3	protein.targeting.chloroplast
M192427	3	44	1st	29.3.4.1	protein.targeting.secretory pathway.ER
G102454	2	1	≥3rd	29.3.4.2	protein.targeting.secretory pathway.golgi
M292815	2	27	2nd	29.3.4.3	protein.targeting.secretory pathway.vacuole

Appendix 1 Continued

Gene ID	K-Means Cluster	Network Degree	Order of neighbors	MapMan (sub)bin	Bin description
M241162	2	35	2nd	29.3.4.3	protein.targeting.secretory pathway.vacuole
G200220	2	11	2nd	29.3.4.3	protein.targeting.secretory pathway.vacuole
M337735	2	22	2nd	29.3.4.3	protein.targeting.secretory pathway.vacuole
M172014	2	15	2nd	29.3.4.3	protein.targeting.secretory pathway.vacuole
M215967	2	3	≥3rd	29.3.4.3	protein.targeting.secretory pathway.vacuole
M293444	2	NA	NA	29.3.4.3	protein.targeting.secretory pathway.vacuole
M261830	2	34	≥3rd	29.3.4.4	protein.targeting.secretory pathway.plasma membrane
M738776	2	NA	NA	29.3.4.4	protein.targeting.secretory pathway.plasma membrane
M149551	2	33	2nd	29.3.4.99	protein.targeting.secretory pathway.unspecified
M344422	2	10	2nd	29.3.4.99	protein.targeting.secretory pathway.unspecified
M914572	3	14	2nd	29.3.4.99	protein.targeting.secretory pathway.unspecified
M332125	3	56	1st	29.3.4.99	protein.targeting.secretory pathway.unspecified
M243164	2	5	≥3rd	29.3.4.99	protein.targeting.secretory pathway.unspecified
M170867	3	9	≥3rd	29.3.4.99	protein.targeting.secretory pathway.unspecified
M265730	2	10	≥3rd	29.3.4.99	protein.targeting.secretory pathway.unspecified
M316688	3	1	≥3rd	29.3.4.99	protein.targeting.secretory pathway.unspecified
M770786	3	2	≥3rd	29.3.4.99	protein.targeting.secretory pathway.unspecified
M231698	3	2	≥3rd	29.3.4.99	protein.targeting.secretory pathway.unspecified
M216604	4	NA	NA	29.3.4.99	protein.targeting.secretory pathway.unspecified
M167586	5	NA	NA	29.3.4.99	protein.targeting.secretory pathway.unspecified
M158475	3	25	2nd	29.4.1.57	protein.postranslational modification.kinase.receptor like cytoplasmatic kinase VII
M262687	3	35	2nd	29.4.1.57	protein.postranslational modification.kinase.receptor like cytoplasmatic kinase VII
M208634	2	21	2nd	29.4.1.57	protein.postranslational modification.kinase.receptor like cytoplasmatic kinase VII
M231519	2	29	2nd	29.4.1.57	protein.postranslational modification.kinase.receptor like cytoplasmatic kinase VII
M307989	3	22	2nd	29.4.1.57	protein.postranslational modification.kinase.receptor like cytoplasmatic kinase VII
M207142	2	15	2nd	29.4.1.57	protein.postranslational modification.kinase.receptor like cytoplasmatic kinase VII
M132677	2	27	2nd	29.4.1.57	protein.postranslational modification.kinase.receptor like cytoplasmatic kinase VII
M273596	2	10	2nd	29.4.1.57	protein.postranslational modification.kinase.receptor like cytoplasmatic kinase VII

Appendix 1 Continued

Gene ID	K-Means Cluster	Network Degree	Order of neighbors	MapMan (sub)bin	Bin description
M139809	2	3	≥3rd	29.4.1.57	protein.postranslational modification.kinase.receptor like cytoplasmatic kinase VII
M310493	4	3	≥3rd	29.4.1.57	protein.postranslational modification.kinase.receptor like cytoplasmatic kinase VII
M314346	4	2	≥3rd	29.4.1.57	protein.postranslational modification.kinase.receptor like cytoplasmatic kinase VII
M870778	5	1	≥3rd	29.4.1.57	protein.postranslational modification.kinase.receptor like cytoplasmatic kinase VII
M287083	5	7	≥3rd	29.4.1.57	protein.postranslational modification.kinase.receptor like cytoplasmatic kinase VII
M912917	5	2	≥3rd	29.4.1.57	protein.postranslational modification.kinase.receptor like cytoplasmatic kinase VII
M695032	6	1	≥3rd	29.4.1.57	protein.postranslational modification.kinase.receptor like cytoplasmatic kinase VII
M274017	6	3	≥3rd	29.4.1.57	protein.postranslational modification.kinase.receptor like cytoplasmatic kinase VII
M266029	6	2	≥3rd	29.4.1.57	protein.postranslational modification.kinase.receptor like cytoplasmatic kinase VII
M148501	6	1	≥3rd	29.4.1.57	protein.postranslational modification.kinase.receptor like cytoplasmatic kinase VII
M230480	6	6	≥3rd	29.4.1.57	protein.postranslational modification.kinase.receptor like cytoplasmatic kinase VII
M526206	6	3	≥3rd	29.4.1.57	protein.postranslational modification.kinase.receptor like cytoplasmatic kinase VII
M712450	2	NA	NA	29.4.1.57	protein.postranslational modification.kinase.receptor like cytoplasmatic kinase VII
M279648	2	NA	NA	29.4.1.57	protein.postranslational modification.kinase.receptor like cytoplasmatic kinase VII
M247000	2	NA	NA	29.4.1.57	protein.postranslational modification.kinase.receptor like cytoplasmatic kinase VII
M599856	3	NA	NA	29.4.1.57	protein.postranslational modification.kinase.receptor like cytoplasmatic kinase VII
M203090	4	NA	NA	29.4.1.57	protein.postranslational modification.kinase.receptor like cytoplasmatic kinase VII
M746889	5	NA	NA	29.4.1.57	protein.postranslational modification.kinase.receptor like cytoplasmatic kinase VII
M203534	5	NA	NA	29.4.1.57	protein.postranslational modification.kinase.receptor like cytoplasmatic kinase VII
M764121	2	15	2nd	29.5.3	protein.degradation.cysteine protease
M813713	3	22	2nd	29.5.3	protein.degradation.cysteine protease
M213404	4	17	2nd	29.5.3	protein.degradation.cysteine protease
M943270	3	15	2nd	29.5.3	protein.degradation.cysteine protease
M197825	3	12	2nd	29.5.3	protein.degradation.cysteine protease
G200718	2	3	2nd	29.5.3	protein.degradation.cysteine protease
M203884	2	11	2nd	29.5.3	protein.degradation.cysteine protease
M232426	2	28	2nd	29.5.3	protein.degradation.cysteine protease
M153556	2	14	≥3rd	29.5.3	protein.degradation.cysteine protease

Appendix 1 Continued

Gene ID	K-Means Cluster	Network Degree	Order of neighbors	MapMan (sub)bin	Bin description
M231666	2	8	≥3rd	29.5.3	protein.degradation.cysteine protease
M185769	2	17	≥3rd	29.5.3	protein.degradation.cysteine protease
M128517	2	4	≥3rd	29.5.3	protein.degradation.cysteine protease
M218404	2	13	≥3rd	29.5.3	protein.degradation.cysteine protease
M242622	2	12	≥3rd	29.5.3	protein.degradation.cysteine protease
M270418	2	12	≥3rd	29.5.3	protein.degradation.cysteine protease
M275507	2	3	≥3rd	29.5.3	protein.degradation.cysteine protease
M163567	6	1	NA	29.5.3	protein.degradation.cysteine protease
M915743	2	NA	NA	29.5.3	protein.degradation.cysteine protease
M287242	5	NA	NA	29.5.3	protein.degradation.cysteine protease
M264884	4	14	2nd	30.2.11	signalling.receptor kinases.leucine rich repeat XI
M585188	3	19	2nd	30.2.11	signalling.receptor kinases.leucine rich repeat XI
M261548	3	27	2nd	30.2.11	signalling.receptor kinases.leucine rich repeat XI
G101113	4	5	2nd	30.2.11	signalling.receptor kinases.leucine rich repeat XI
M144320	3	24	2nd	30.2.11	signalling.receptor kinases.leucine rich repeat XI
M934381	4	28	2nd	30.2.11	signalling.receptor kinases.leucine rich repeat XI
M798156	3	50	1st	30.2.11	signalling.receptor kinases.leucine rich repeat XI
M897253	4	45	1st	30.2.11	signalling.receptor kinases.leucine rich repeat XI
M186555	3	41	1st	30.2.11	signalling.receptor kinases.leucine rich repeat XI
M322989	2	18	≥3rd	30.2.11	signalling.receptor kinases.leucine rich repeat XI
M162064	5	1	≥3rd	30.2.11	signalling.receptor kinases.leucine rich repeat XI
M315498	5	5	≥3rd	30.2.11	signalling.receptor kinases.leucine rich repeat XI
M266980	5	5	≥3rd	30.2.11	signalling.receptor kinases.leucine rich repeat XI
M188950	5	2	≥3rd	30.2.11	signalling.receptor kinases.leucine rich repeat XI
M747845	5	1	≥3rd	30.2.11	signalling.receptor kinases.leucine rich repeat XI
M304719	6	1	≥3rd	30.2.11	signalling.receptor kinases.leucine rich repeat XI
M523939	6	2	≥3rd	30.2.11	signalling.receptor kinases.leucine rich repeat XI
M812673	6	4	≥3rd	30.2.11	signalling.receptor kinases.leucine rich repeat XI

Appendix 1 Continued

Gene ID	K-Means Cluster	Network Degree	Order of neighbors	MapMan (sub)bin	Bin description
M167358	1	NA	NA	30.2.11	signalling.receptor kinases.leucine rich repeat XI
M471111	1	NA	NA	30.2.11	signalling.receptor kinases.leucine rich repeat XI
M525602	2	NA	NA	30.2.11	signalling.receptor kinases.leucine rich repeat XI
M129456	5	NA	NA	30.2.11	signalling.receptor kinases.leucine rich repeat XI
M140341	6	NA	NA	30.2.11	signalling.receptor kinases.leucine rich repeat XI
M233031	3	9	2nd	30.2.17	signalling.receptor kinases.DUF 26
M139111	4	10	2nd	30.2.17	signalling.receptor kinases.DUF 26
M241508	4	12	2nd	30.2.17	signalling.receptor kinases.DUF 26
M157044	4	40	1st	30.2.17	signalling.receptor kinases.DUF 26
M124621	2	8	≥3rd	30.2.17	signalling.receptor kinases.DUF 26
M086517	3	5	≥3rd	30.2.17	signalling.receptor kinases.DUF 26
M921241	3	1	≥3rd	30.2.17	signalling.receptor kinases.DUF 26
M205882	5	2	≥3rd	30.2.17	signalling.receptor kinases.DUF 26
M205079	1	NA	NA	30.2.17	signalling.receptor kinases.DUF 26
G107157	3	NA	NA	30.2.17	signalling.receptor kinases.DUF 26
M136671	2	18	2nd	31.1	cell.organisation
G107847	2	16	2nd	31.1	cell.organisation
M572047	2	11	2nd	31.1	cell.organisation
M497445	2	4	2nd	31.1	cell.organisation
M408630	2	1	2nd	31.1	cell.organisation
M247777	2	24	2nd	31.1	cell.organisation
M303061	5	10	2nd	31.1	cell.organisation
M319102	2	18	2nd	31.1	cell.organisation
G303112	2	13	2nd	31.1	cell.organisation
G100980	2	21	2nd	31.1	cell.organisation
M437009	2	22	2nd	31.1	cell.organisation
M774288	2	21	2nd	31.1	cell.organisation
M382436	2	56	1st	31.1	cell.organisation

Appendix 1 Continued

Gene ID	K-Means Cluster	Network Degree	Order of neighbors	MapMan (sub)bin	Bin description
M477969	2	48	1st	31.1	cell.organisation
M691891	2	5	≥3rd	31.1	cell.organisation
M137341	2	5	≥3rd	31.1	cell.organisation
M474142	2	2	≥3rd	31.1	cell.organisation
M248977	3	6	≥3rd	31.1	cell.organisation
M345477	3	7	≥3rd	31.1	cell.organisation
M752428	1	2	≥3rd	31.1	cell.organisation
M756723	3	2	≥3rd	31.1	cell.organisation
G104205	3	2	≥3rd	31.1	cell.organisation
M222020	5	1	≥3rd	31.1	cell.organisation
G105192	5	5	≥3rd	31.1	cell.organisation
M912745	5	1	≥3rd	31.1	cell.organisation
M225059	2	1	NA	31.1	cell.organisation
M509072	2	1	NA	31.1	cell.organisation
M812416	1	NA	NA	31.1	cell.organisation
M482268	2	NA	NA	31.1	cell.organisation
M465675	2	NA	NA	31.1	cell.organisation
M525934	2	NA	NA	31.1	cell.organisation
M884170	2	NA	NA	31.1	cell.organisation
M292132	2	NA	NA	31.1	cell.organisation
M745595	2	NA	NA	31.1	cell.organisation
M213737	2	NA	NA	31.1	cell.organisation
M776374	2	NA	NA	31.1	cell.organisation
M152676	5	NA	NA	31.1	cell.organisation
M749824	6	NA	NA	31.1	cell.organisation
G103492	2	35	2nd	31.1.1.3.11	cell.organisation.cytoskeleton.Myosin.Class XI
M252114 (Ma1)	6	NA	NA	34.8.1	transport.malate transporters at the envelope membran

* The letter ‘M’ in gene IDs is abbreviated from ‘MDP0000’ in the original apple gene IDs used at the Genome Database for Rosaceae. Genes starting with ‘G’ are novel transcripts in the improved apple reference transcriptome (Bai et al. 2014)

Appendix 2 The list of 303 genes associated with *Ma1* and expressed differentially between *mama* and *Ma__* groups

Gene ID**	RPKM of *mama*	RPKM of *Ma__*	Network Degree	Order of *Ma1* neighbors *	MapMan (sub)bin	Gene annotation
M235175	7.31	22.03	NA	NA	1.1.1.1	Chlorophyll a-b binding protein 3, chloroplastic, photosystem II, LHC-II
M309425	3.77	14.36	10	2nd	1.1.1.1	Chlorophyll a-b binding protein 3, chloroplastic, photosystem II, LHC-II
M293052	1.17	6.79	12	2nd	1.1.1.1	Chlorophyll a-b binding protein 3, chloroplastic, photosystem II, LHC-II
M364491	16.62	30.60	3	≥3rd	1.1.3	Cytochrome b6f complex subunit, light reaction
M800352	2.93	13.00	10	≥3rd	1.1.1.2	Photosystem II 10 kDa polypeptide
M322896	4.52	10.63	3	≥3rd	1.3.12	PRK, Phosphoribulokinase//uridine kinase, calvin cycle
M223905	15.60	41.03	4	≥3rd	1.3.13	RuBisCO, ribulose-1 5-bisphosphate carboxylase/oxygenase activase, calvin cycle
M731480	36.56	85.44	6	≥3rd	1.3.2	RuBisCO small subunit, bibulose bisphosphate carboxylase small chain
M811918	2.17	6.06	6	≥3rd	1.1.5.3	Ferredoxin-NADP reductase
M618810	0.85	3.43	10	≥3rd	1.2.4.4	Glycine cleavage system H protein, photorespiration
M273014	1.08	5.33	13	≥3rd	1.3.7	FBPase, Fructose-1,6-bisphosphatase (EC 3.1.3.11), calvin cycle
G107222	7.29	13.49	6	2nd	2.1.2.1	ADP-glucose pyrophosphorylase
M346313	28.65	15.37	3	≥3rd	3.6	Glucan synthase, callose synthesis
M163436	8.99	2.53	9	2nd	3.6	Glucan synthase, callose synthesis
M251480	16.17	5.71	1	≥3rd	4.2.4	PFK, phosphofructokinase, glycolysis
M130296	3.89	9.12	1	≥3rd	6.5	PPDK, Pyruvate phosphate dikinase 1
M191667	6.33	3.00	3	≥3rd	8.1.3	Aconitase (EC 4.2.1.3) in TCA
M682401	0.04	1.65	13	1st	8.1.1.2	TCA, pyruvate DH. Protein dihydrolipoyllysine-residue acetyltransferase component of pyruvate dehydrogenase complex
M163246	96.65	81.87	5	2nd	8.1.7	SDH, succinate dehydrogenase in TCA
M276121	28.94	20.47	12	2nd	9.5	Ubiquinol-cytochrome C reductase iron-sulfur subunit, mitochondrial, ATP synthesis
M448752	67.11	87.54	4	≥3rd	10.2.1	Cellulose synthase/transferase
M259640	2.15	5.12	1	≥3rd	10.7	Expansin, cell wall modification
M212156	7.18	13.92	1	≥3rd *	10.1.6	GAE, UDP-glucuronate 4-epimerase, cell wall precursor synthesis
M904458	292.12	837.41	4	≥3rd	10.5.1.1	cell wall protein FLA, Fasciclin-like arabinogalactan, anchored to plasma membrane
M130769	2.91	5.51	7	≥3rd	10.6.3	Polygalacturonase, cell wall degradation
M573746	64.99	10.89	8	≥3rd	10.6.3	Pectate lyase, cell wall degradation
M581832	3.25	6.99	11	2nd	10.6.3	Polygalacturonase, cell wall degradation

Appendix 2 Continued

Gene ID	RPKM of *mama*	RPKM of *Ma__*	Network Degree	Order of *Ma1* neighbors *	MapMan (sub)bin	Gene annotation
G100104	43.08	23.43	1	≥3rd	11.4	Oleosin family protein in monolayer-surrounded lipid storage body
M234830	11.51	5.99	5	≥3rd	11.5.3	FAD-dependent glycerol-3-phosphate dehydrogenase
M286450	26.03	18.03	NA	NA	11.8.1	Sphingosine hydroxylase, lipid metabolism
M911376	4.34	7.44	5	1st	11.8.10	Phosphatidylcholinesterol O-acyltransferase, lipid metabolism
M760376	37.04	74.48	9	2nd	11.8.1	Fatty acid desaturase, lipid metabolism
M224857	1.91	5.31	11	2nd	11.1.9	Long-chain-fatty-acid CoA ligase, FA synthesis and FA elongation
G105402	17.89	31.37	3	≥3rd	12.3.1	mitochondrial glutamate dehydrogenase.
G202922	8.31	2.11	5	≥3rd	12.2.2	Glutamine synthetase cytosolic isozyme (EC 6.3.1.2), ammonia metabolism
G200531	4.64	1.07	8	≥3rd	12.2.2	Glutamine synthetase cytosolic isozyme (EC 6.3.1.2), ammonia metabolism
M930168	1.67	4.54	1	≥3rd	13.1.4.1	Ketol-acid reductoisomerase, amino acid synthesis
M361922H	10.73	2.51	10	2nd	13.2.2.3	Arginine decarboxylase (ADC1), glutamate family degradation
G202653	17.39	11.15	2	≥3rd	13.1.5.2.41	Sarcosine oxidase, amino acid synthesis
M686885	950.26	532.89	14	2nd	13.1.5.2.41	Sarcosine oxidase, amino acid synthesis
M593587	48.50	35.22	5	2nd	13.1.6.5.5	Tryptophan synthase beta chain, aromatic aa synthesis
M136894	43.88	30.98	1	≥3rd	16.2	Putative aminotransferase
M261201	20.13	39.29	8	≥3rd	16.1.1	Geranylgeranyl reductase, secondary metabolism, isoprenoids
G200494	62.92	32.31	5	≥3rd	16.1.3.1	Hydroxyphenylpyruvate dioxygenase (HPPDase), secondary metabolism, isoprenoids, tocopherol biosynthesis
M138071	7.76	1.69	7	2nd	16.1.2.2	Hydroxymethylglutaryl-CoA synthase, secondary metabolism, isoprenoids, mevalonate pathway
M268909	5.38	9.62	2	≥3rd	16.1.2.3	3-hydroxy-3-methylglutaryl-coenzyme A reductase, , secondary metabolism, isoprenoids, mevalonate pathway
M242980	4.91	9.42	3	≥3rd	16.1.5	Cycloartenol synthase, secondary metabolism, isoprenoids, terpenoids
M241703	14.88	26.05	4	≥3rd	16.1.4.2	Phytoene desaturase, secondary metabolism, isoprenoids, carotenoids
M148978	7.05	12.33	7	2nd	16.1.4.2	Phytoene desaturase, secondary metabolism, isoprenoids, carotenoids
M157209	49.81	27.23	10	≥3rd	16.8.3.1	Cinnamoyl CoA reductase-like protein, secondary metabolism, flavonoids
M239946	105.45	69.61	3	≥3rd	16.8.5.1	Isoflavone reductase related protein, secondary metabolism, flavonoids
M141719	2.12	13.00	9	≥3rd	17.2.3	Auxin-responsive SAUR-like protein
M222749	3.34	6.94	9	2nd	17.2.3	Auxin-induced protein

Appendix 2 Continued

Gene ID	RPKM of *mama*	RPKM of *Ma__*	Network Degree	Order of *Ma1* neighbors *	MapMan (sub)bin	Gene annotation
M747141	1.30	3.45	4	≥3rd	17.4.2	Histidine kinase, cytokinin signal transduction
M747140	1.60	5.33	6	≥3rd	17.4.2	Histidine kinase, cytokinin signal transduction
G100505	7.34	3.72	9	≥3rd	17.3.2.1	LRR Receptor Kinase; BRI1 like (BRL1), brassinosteroid signal transduction, in plasma membrane
M195089	0.12	1.31	2	≥3rd	18.7	Quinolinate synthetase, 4 iron, 4 sulfur cluster binding
M287302	14.51	2.82	NA	NA	20.1	Thaumatin-like protein
M789304	7.82	1.00	3	≥3rd	20.1	Acidic endochitinase
M670775	12.95	6.06	4	2nd	20.1	avirulence-responsive protein
G107736	1.45	4.41	4	2nd	20.1	LRR and NB-ARC domains-containing disease resistance protein
M272762	2.67	6.74	7	2nd	20.1	LRR disease resistance protein
M439929	22.62	15.61	2	≥3rd	20.2.1	DNAJ chaperoneN-terminal domain-containing protein
M295157	3.99	7.18	2	≥3rd	20.2.1	Heat shock 70 kDa protein
M213035	0.18	1.53	6	≥3rd	20.2.1	DNAJ chaperone C-terminal domain-containing protein
M632148	35.93	22.42	5	≥3rd	20.2.2	Cold-shock DNA-binding family protein
M299010	2.51	9.51	2	≥3rd	20.2.3	dehydration-responsive family protein
M148807	10.84	16.06	16	2nd	20.2.3	Methyltransferase
M293686	1.78	3.91	5	≥3rd	20.1.7	Disease resistance protein
M758643	5.38	23.03	6	2nd	20.1.7	TIR-NBS-LRR disease resistance protein
M865880	9.84	18.13	10	2nd	20.1.7	COP1-interacting protein-like, disease resistance
M212150	7.46	13.45	3	≥3rd	21.1	Thioredoxin
M828993	34.80	24.16	5	≥3rd	21.1	Thioredoxin
G104166	1.16	3.99	4	≥3rd	20.2.99	RmlC-like cupins superfamily protein, manganese ion binding
M894544	110.04	81.61	14	2nd	20.2.99	Ozone-responsive stress related protein
M896377	142.39	115.59	NA	NA	21.2	Putative progesterone receptor membrane component 1
M149697	6.29	1.38	NA	NA	26.1	Cytochrome P450
M701561	5.92	2.09	8	2nd	26.1	Cytochrome P450
M231959	118.80	62.41	11	2nd	26.1	Epoxide hydrolase 2
M213291	65.91	159.32	NA	NA	26.2	Glycosyltransferase 5

Appendix 2 Continued

Gene ID	RPKM of *mama*	RPKM of *Ma__*	Network Degree	Order of *Ma1* neighbors *	MapMan (sub)bin	Gene annotation
M262871	1.51	0.02	6	≥3rd	26.2	Indole-3-acetate beta-glucosyltransferase (EC 2.4.1.121), UDP-glycosyltransferase
M803674	9.12	17.98	12	2nd	26.2	Indole-3-acetate beta-glucosyltransferase (EC 2.4.1.121), UDP-glycosyltransferase
M752561	4.66	0.32	19	1st	26.2	Indole-3-acetate beta-glucosyltransferase (EC 2.4.1.121), UDP-glycosyltransferase
M351526	0.31	2.76	5	≥3rd	26.21	protease inhibitor/seed storage/lipid transfer protein (LTP) family protein
M760940	0.20	1.28	2	2nd	26.22	short-chain dehydrogenase/reductase (SDR) family protein
M265006	14.28	5.53	4	≥3rd	26.22	Dehydrogenase/reductase SDR family protein
M369466	1.32	4.17	2	≥3rd *	26.3	ER degradation-enhancing alpha-mannosidase-like protein
M759437	4.49	8.91	8	≥3rd	26.3	Rieske (2Fe-2S) domain protein
M299794	1.22	3.50	9	≥3rd	26.4	Glucan endo-1,3-beta-glucosidase
M146967	1.14	3.12	8	≥3rd	26.4.1	Putative glucan endo-1-3-beta-glucosidase
M355032	6.91	10.96	5	≥3rd	26.8	Putative methylesterase 14, chloroplastic
M511650	18.11	2.75	4	≥3rd	26.9	Glutathione S-transferase
M170114	2.18	6.56	1	≥3rd	27.1.1	U1 small nuclear ribonucleoprotein A
M570413	12.47	1.00	6	≥3rd	27.3.12	Zinc finger CCCH domain-containing protein
M190273	20.75	63.12	5	2nd	27.3.19	Ethylene-insensitive 3 like (EIL)
M290818	7.75	13.88	8	2nd	27.3.20	Aribidopsis response regulator 1 (ARR1)
M224740	3.93	7.30	9	2nd	27.3.20	Aribidopsis response regulator 1 (ARR1)
M950387	1.48	4.66	2	≥3rd	27.3.21	DELLA protein GAI (Gibberellic acid-insensitive protein 1)
G303417	3.04	7.55	5	≥3rd	27.3.22	HB, homeobox transcription factor, response to sucrose stimulus, locate in nuleus
M509120	281.73	180.37	5	≥3rd	27.3.22	Homeobox-leucine zipper protein
M239684	64.32	40.26	9	≥3rd	27.3.22	Homeobox protein BLH6 (BELL1-like homeodomain 6)
M423596	6.53	15.80	17	2nd	27.3.22	Homeobox-leucine zipper protein, ATHB13
M190504	10.11	4.86	4	≥3rd	27.3.3	Ethylene-responsive transcription factor
G105310	4.02	1.28	9	2nd	27.3.3	Ethylene-responsive transcription factor
M136037	26.16	50.62	2	≥3rd *	27.3.37	LOB domain-containing protein
M945260	38.71	69.11	10	2nd	27.3.40	Auxin-responsive transcription factor
G102193	68.76	53.76	2	≥3rd	28.1	ATP-dependent DEAD-Box RNA helicase, invoved in poly(A)+ mRNA export from nucleus

Appendix 2 Continued

Gene ID	RPKM of *mama*	RPKM of *Ma__*	Network Degree	Order of *Ma1* neighbors *	MapMan (sub)bin	Gene annotation
M139525	1.23	4.28	7	≥3rd	28.2	DNA-3-methyladenine glycosylase, putative
M083546	0.45	4.01	2	≥3rd	27.3.99	unclassified transcription factor
M157639	0.11	2.61	2	≥3rd *	27.3.99	unclassified transcription factor, putative CCCH zinc finger factor
G106350	5.85	13.61	3	≥3rd	27.3.99	unclassified transcription factor
G104358	70.34	119.26	1	≥3rd	29.3.1	importin alpha 1, protein transporter, locate in nucleus
M124554	7.16	2.39	2	2nd	29.4	ankyrin protein kinase
M291428	6.71	2.60	2	2nd	29.4	Protein serine/threonine/tyrosine kinase
G104359	31.05	67.33	3	≥3rd	29.3.1	importin alpha 3, protein transporter, locate in nucleus
M437033	11.09	4.33	3	≥3rd	29.4	Protein phosphatase 2C
M628976	45.20	14.72	3	≥3rd	29.4	Protein phosphatase 2C
M304124	49.47	15.60	6	≥3rd	29.4	Protein phosphatase 2C
G104167	1.35	13.63	7	2nd	29.4	tyrosine/serine/threonine phosphatase
G103047	0.32	1.90	8	≥3rd	29.4	Serine/threonine phosphatase
M251050	4.89	2.07	9	≥3rd	29.2.2	60S ribosome subunit biogenesis protein NIP7
M250124	4.68	9.63	20	1st	29.4	Serine/threonine protein kinase
G300046	10.73	4.62	2	≥3rd	29.2.2.1	General control non-repressible 1 (GCN 1), ribosome biogenesis, export from nucleus
M189832	4.50	11.74	5	2nd	29.5	Metalloendopeptidase, protein degradation
M168289	18.21	11.95	8	≥3rd	29.5	Metalloendopeptidase, protein degradation
M382190	44.61	56.86	9	≥3rd	29.5	Metacaspase 1, protein degradation
M219184	133.32	94.84	NA	NA	29.5.1	Subtilisin-like protease 2
M239581	7.42	11.43	12	2nd	29.7	Sialyltransferase, glycosyl transferase family protein
M303249	42.85	28.88	2	≥3rd	29.2.1.2.2.10	60S ribosomal protein L10
M294334	0.22	1.52	NA	NA	29.5.11.20	26S proteasome non-ATPase regulatory subunit
G103136	1854.22	1229.56	6	2nd	29.5.4	asparate proteinase stable over a broad pH range (ph 3-8), endopeptidase activity, response to salt stress, in vacuole
M654592	9.61	17.20	2	2nd	29.5.5	Serine protease, chloroplastic
M193152	14.66	20.26	3	≥3rd	29.5.5	Tripeptidyl peptidase II, protein degradation
M792008	10.79	18.21	4	≥3rd	29.5.5	Serine protease, chloroplastic

Appendix 2 Continued

Gene ID	RPKM of *mama*	RPKM of *Ma__*	Network Degree	Order of *Ma1* neighbors *	MapMan (sub)bin	Gene annotation
M906581	1.65	3.88	1	≥3rd	29.4.1.57	Receptor kinase cytoplastic
M148287	10.57	20.24	5	2nd	29.4.1.57	Receptor kinase cytoplastic
M145521	4.40	12.51	6	≥3rd	29.4.1.57	Serine/threonine receptor kinase cytoplastic
M163242	1.95	4.12	6	2nd	29.4.1.57	Serine/threonine receptor kinase cytoplastic
M545129	7.36	17.14	8	2nd	29.4.1.57	Receptor kinase cytoplastic
M151537	18.66	33.38	8	2nd	29.4.1.57	Receptor kinase cytoplastic
M651862	2.21	5.76	13	2nd	29.4.1.57	BAK1, an LRR Receptor-like protein kinase, binding BRI1 and modulates brassinosteroid signaling
M788186	9.31	13.86	6	≥3rd	30.1	Glutamate receptor, sugar and nutrient signalling
M271605	5.09	11.23	NA	NA	30.11	BTB/POZ domain-containing protein
M163893	10.17	14.81	NA	NA	30.11	Protein TIC, light signaling
G104596	11.74	30.93	11	2nd	30.11	Phototropic-responsive NPH3 family protein, light signalling
M307855	2.68	8.99	13	2nd	30.11	Phototropism protein, light signalling
M434420	2.42	0.57	3	≥3rd	29.5.11.4.2	Ring finger protein, ubiquitin
G103529	10.67	16.28	3	≥3rd	30.3	Ca2+-ATPase with an N-terminal autoinhibitor (ACA1), in chloroplast inner membrane
M140330	2.94	6.78	13	1st	30.3	Calmodulin-like protein, calcium singalling
M841118	1.35	3.27	14	2nd	30.3	Calcium-binding protein
M319170	9.15	15.14	25	1st	30.3	IQ domain-containing protein, calmodulin binding
M138674	0.08	1.09	1	≥3rd *	30.2.11	Leucine rich repeat receptor kinase
M525602	0.08	14.74	7	≥3rd	30.2.11	Leucine rich repeat receptor kinase
M138489	8.60	13.82	11	2nd	30.2.11	Leucine rich repeat receptor kinase
M143570	0.30	2.64	2	≥3rd	30.2.17	Serine/threonine receptor kinase
M311146	22.87	32.93	4	2nd	30.5	Ras-related protein Rab11D, G protein signalling
G104738	19.43	12.92	7	2nd	30.5	Ras-related small GTP-binding family protein
M843303	77.68	130.52	2	≥3rd	29.5.11.4.3.2	Kelch-like Fbox protein, ubiquitin
M319818	2.15	4.42	21	2nd	29.5.11.4.3.2	F-box protein, ubiquitin
M344422	10.79	15.95	2	≥3rd	29.3.4.99	ADP-ribosylation factor
G107229	2.97	5.74	NA	NA	31.1	P-loop containing nucleoside triphosphate hydrolases superfamily protein, microtulin

Appendix 2 Continued

Gene ID	RPKM of *mama*	RPKM of *Ma__*	Network Degree	Order of *Ma1* neighbors *	MapMan (sub)bin	Gene annotation
M509076	2.90	5.64	8	≥3rd	31.1	Ankyrin repeat-containing protein, putative
M163222	3.72	10.14	14	2nd	31.1	Fimbrin, organisation
M168542	12.88	28.80	7	2nd	30.2.99	Leucine rich repeat receptor kinase
M407802	16.71	10.69	5	≥3rd	31.4	Exocyst complex component, vasicle transport
G101937	8.86	14.35	16	2nd	31.4	SNARE-like superfamily protein, involved in vesicle-mediated transport, locate in plasma membrane
M850643	195.71	401.05	2	≥3rd	33.2	late embryogenesis abundant group 1 domain-containing protein
G107922	0.42	1.94	1	≥3rd	33.99	EamA-like transporter family protein, development
M120881	27.83	56.06	2	≥3rd	33.99	NAC domain-containing protein, developmental and stress responsive transcription factors
M142574	8.03	18.54	2	≥3rd *	33.99	Tetraspanin 2 (TET2), senescence-associated protein
M170822	21.19	35.76	4	≥3rd	33.99	Tetraspanin 8 (TET8), senescence-associated protein
M126517	15.50	4.23	12	≥3rd	33.99	NAC domain-containing protein, developmental and stress responsive transcription factors
G105925	22.12	14.07	13	2nd	33.99	Protein with RING/U-box and TRAF-like domains, ubiquitin-protein ligase, zinc ion binding
M269759	45.87	18.47	NA	NA	34.13	Nitrate transporter
M293045	1.23	9.11	1	≥3rd	34.13	Proton-dependent oligopeptide transport
M322339	0.15	1.79	1	≥3rd	34.15	Voltage-gated potassium channe
M231597	12.09	23.18	2	≥3rd	34.16	ABC transporter family protein, multidrug resistance protein
M269936	6.59	10.20	2	2nd	34.16	ABC transporter family protein, multidrug resistance protein
M942052	2.00	4.40	3	≥3rd	34.16	ABC transporter family protein, pleiotropic drug resistance protein
M142911	11.45	20.75	8	≥3rd	34.18	Anion exchange protein. Boron transporter (dicarboxylate)
M296050	11.70	5.62	2	≥3rd	34.2	Polyol transporter, sugar and organic acid transporter
M940086	3.85	1.54	6	≥3rd	34.2	Polyol transporter, sugar and organic acid transporter
M834327	3.18	18.27	7	2nd	34.22	Cyclic nucleotide gated channel, calmodulin binding
G103588	47.68	259.80	7	2nd	34.19.1	Aquaporin,plasma membrane intrinsic protein 2 (PIP2)
M134278	5.40	1.89	5	2nd	34.3	Amino acid permease family protein
M246926	12.88	7.51	8	≥3rd	34.19.4	Aquaporin, node 26 like intrinsic protein (NIP)
M252114	11.70	19.62	13	1st	34.8.1	Aluminum activated malate transporter

Appendix 2 Continued

Gene ID	RPKM of *mama*	RPKM of *Ma__*	Network Degree	Order of *Ma1* neighbors *	MapMan (sub)bin	Gene annotation
M219689	59.70	45.25	NA	NA	35.1	5'-AMP-activated protein kinase subunit gamma-3
M138424	3.92	1.33	NA	NA	35.1	Methyltransferase-like protein
M299828	43.35	28.76	1	≥3rd	35.1	Octicosapeptide/Phox/Bem1p domain-containing protein
M737765	1.58	3.84	2	≥3rd	35.1	Macrophage migration inhibitory factor-like protein
M779889	49.28	36.62	2	≥3rd	35.1	Putative integral membrane protein
M127844	3.78	8.98	2	≥3rd	35.1	Sigma factor sigb regulation protein rsbq, putative
M364253	23.05	5.33	3	≥3rd	35.1	CBS domain-containing protein
M282814	11.89	6.40	3	≥3rd	35.1	Putative metal ion-binding protein
G104393	1.20	3.37	4	≥3rd	35.1	unknown protein containing retrotransposon gag protein
M174738	0.39	2.32	4	2nd	35.1	Leucine-rich repeat-containing protein, putative
M191472	0.59	3.03	NA	NA	35.2	DEAD box ATP-dependent RNA helicase, putative
G106731	0.22	1.77	NA	NA	35.2	Unknown protein
M283400	1.30	7.54	NA	NA	35.2	Unknown protein
G100470	7.06	33.50	NA	NA	35.2	Unknown protein
G107103	3.12	9.02	NA	NA	35.2	Unknown protein
G100416	13.25	25.25	NA	NA	35.2	Unknown protein
G101209	8.85	13.35	NA	NA	35.2	Unknown protein
G200112	54.85	79.51	NA	NA	35.2	Unknown protein
M125604	38.21	19.02	NA	NA	35.2	Unknown protein
M783350	11.52	3.33	NA	NA	35.2	Unknown protein
M134543	12.62	6.11	1	≥3rd	35.2	CASP-like protein
G102318	0.12	1.60	1	≥3rd	35.2	Unknown protein
M369115	0.12	1.10	1	≥3rd	35.2	Unknown protein
G104029	0.41	3.66	1	≥3rd	35.2	Unknown protein
G102465	1.16	3.61	1	≥3rd	35.2	Unknown protein
G107006	2.96	6.19	1	≥3rd	35.2	Unknown protein
G103716	15.27	24.39	1	≥3rd	35.2	Unknown protein
G200125	12.05	6.14	1	≥3rd	35.2	Unknown protein

Appendix 2 Continued

Gene ID	RPKM of *mama*	RPKM of *Ma__*	Network Degree	Order of *Ma1* neighbors *	MapMan (sub)bin	Gene annotation
M868659	1.14	4.61	1	≥3rd *	35.2	AT5G28150-like protein
M252944	2.80	6.24	1	≥3rd *	35.2	Putative pentatricopeptide repeat protein
G100038	0.03	3.69	1	≥3rd *	35.2	Unknown protein
G100039	0.29	2.52	1	≥3rd *	35.2	Unknown protein
G108003	19.28	63.07	1	≥3rd *	35.2	Unknown protein
M477255	1.93	5.25	1	≥3rd *	35.2	Unknown protein
M123003	11.04	20.79	2	≥3rd	35.2	AT5g17280/MKP11_13
M234882	0.89	4.92	2	≥3rd	35.2	Integrator complex subunit 9
M240442	12.87	5.16	2	≥3rd	35.2	Putative uncharacterized protein 5K14.3
M598822	15.85	40.51	2	≥3rd	35.2	Ring-infected erythrocyte surface antigen
M352138	0.82	2.87	2	≥3rd	35.2	Unknown protein
G104788	3.72	10.97	2	≥3rd	35.2	Unknown protein
G106639	3.60	7.67	2	≥3rd	35.2	Unknown protein
G107554	12.20	20.86	2	≥3rd	35.2	Unknown protein
G101368	12.77	21.43	2	≥3rd	35.2	Unknown protein
M139303	19.69	9.66	2	≥3rd	35.2	Unknown protein
G107650	10.38	4.57	2	≥3rd	35.2	Unknown protein
M326720	4.64	1.91	2	≥3rd	35.2	Unknown protein
G100435	0.13	1.72	2	≥3rd *	35.2	Unknown protein
M214283	0.55	2.10	2	≥3rd *	35.2	Unknown protein
M453055	2.61	5.86	3	≥3rd	35.2	At1g29190/F28N24_12
M219042	6.31	18.61	3	≥3rd	35.2	Unknown protein
G101968	8.52	17.95	3	≥3rd	35.2	Unknown protein
M884644	6.18	11.37	3	≥3rd	35.2	Unknown protein
M872691	22.39	32.92	3	≥3rd	35.2	Unknown protein
G101851	4.59	1.80	3	≥3rd	35.2	Unknown protein
M308221	10.97	15.56	3	2nd	35.2	Emb|CAB79781.1
M143488	22.75	42.88	4	≥3rd	35.2	uncharacterized Cys-rich protein

Appendix 2 Continued

Gene ID	RPKM of *mama*	RPKM of *Ma__*	Network Degree	Order of *Ma1* neighbors *	MapMan (sub)bin	Gene annotation
G104762	0.10	1.02	4	≥3rd	35.2	Unknown protein
G103277	0.42	4.02	4	≥3rd	35.2	Unknown protein
M357045	0.17	1.49	4	≥3rd	35.2	Unknown protein
M349063	1.17	5.15	4	≥3rd	35.2	Unknown protein
M328020	3.56	8.66	4	≥3rd	35.2	Unknown protein
M347092	3.65	8.38	4	≥3rd	35.2	Unknown protein
M367636	19.75	43.34	4	≥3rd	35.2	Unknown protein
M246085	3.55	6.90	4	≥3rd	35.2	Unknown protein
G107007	10.75	19.77	4	≥3rd	35.2	Unknown protein
G108038	7.70	12.14	4	≥3rd	35.2	Unknown protein
G103142	0.90	4.22	4	≥3rd *	35.2	Unknown protein
G103896	2.46	6.10	4	2nd	35.2	Unknown protein
M885511	4.62	1.33	5	≥3rd	35.2	Sigma factor binding protein 1
M368785	0.60	2.01	5	≥3rd	35.2	Unknown protein
G104304	8.10	17.98	5	≥3rd	35.2	Unknown protein
G105620	4.18	7.40	5	≥3rd	35.2	Unknown protein
G103578	13.65	22.48	5	≥3rd	35.2	Unknown protein
G107777	0.06	1.41	5	2nd	35.2	Unknown protein
G106725	18.02	28.01	5	2nd	35.2	Unknown protein
G103283	1.96	11.60	6	1st	35.2	Unknown protein
M124546	2.33	8.59	6	≥3rd	35.2	CASP-like protein
M592961	1.40	5.68	6	≥3rd	35.2	Ice binding protein, putative
M191749	83.46	156.67	6	≥3rd	35.2	Neurofilament medium polypeptide
G104726	8.19	13.78	6	≥3rd	35.2	Unknown protein
G100682	3.73	0.67	6	≥3rd	35.2	Unknown protein
G200162	17.54	10.41	6	≥3rd	35.2	Unknown protein in plastid
M680536	6.99	3.27	6	2nd	35.2	LSM14-like protein
M352691	5.68	9.35	6	2nd	35.2	Unknown protein

Appendix 2 Continued

Gene ID	RPKM of *mama*	RPKM of *Ma__*	Network Degree	Order of *Ma1* neighbors *	MapMan (sub)bin	Gene annotation
M745662	6.29	2.88	6	2nd	35.2	Unknown protein
M139221	7.23	16.95	7	≥3rd	35.2	IMP dehydrogenase/GMP reductase, putative
G104764	0.17	3.80	7	≥3rd	35.2	Unknown protein
G104741	2.80	15.79	7	≥3rd	35.2	Unknown protein
M351809	12.29	21.51	7	≥3rd	35.2	Unknown protein
M575203	12.37	21.57	7	2nd	35.2	Putative uncharacterized protein B1168F12.31
G100927	0.04	0.96	7	2nd	35.2	Unknown protein
M744636	2.95	0.62	8	1st	35.2	Polyphenol oxidase (EC 1.10.3.1), chloroplastic
G105483	3.25	7.65	8	1st	35.2	Unknown protein
M143636	3.85	7.39	8	≥3rd	35.2	Carboxyl-terminal proteinase
M163533	2.39	0.38	8	≥3rd	35.2	Unknown protein
M234782	7.78	1.45	8	2nd	35.2	Polyphenol oxidase (EC 1.10.3.1), chloroplastic
G103284	2.11	8.56	8	2nd	35.2	Unknown protein
M442350	0.60	5.46	9	1st	35.2	Unknown protein
M278260	210.51	374.73	9	2nd	35.2	Putative uncharacterized protein P0022B05.126
G105841	20.98	39.47	9	2nd	35.2	Unknown protein
M122554	3.09	6.59	10	1st	35.2	Putative uncharacterized protein T2E22.10
M716241	0.55	3.03	10	≥3rd	35.2	AT3g50680/T3A5_60
M352743	6.97	12.65	10	≥3rd	35.2	Unknown protein
G200575	120.87	88.08	11	1st	35.2	Unknown protein
M369858	5.61	9.29	11	≥3rd	35.2	dCTP pyrophosphatase
M306372	9.97	14.98	12	≥3rd	35.2	Putative uncharacterized protein (Precursor)
G107102	0.94	10.32	12	≥3rd	35.2	Unknown protein
G102909	0.47	2.90	12	2nd	35.2	Unknown protein
M202406	2.05	4.89	13	2nd	35.2	Desumoylating isopeptidase 2
G101928	18.67	30.65	13	2nd	35.2	Unknown protein
G203972	6.54	20.37	16	2nd	35.2	Unknown protein
M233503	15.90	31.34	17	2nd	35.2	Putative uncharacterized protein RAF9-1

Appendix 2 Continued

Gene ID	RPKM of *mama*	RPKM of *Ma__*	Network Degree	Order of *Ma1* neighbors *	MapMan (sub)bin	Gene annotation
M183277	16.46	31.61	17	1st	35.2	Putative uncharacterized protein T21H19_30
M242522	50.27	60.66	3	≥3rd	35.1.19	Elicitor responsive protein 3
M564732	1.83	5.80	5	≥3rd	35.1.3	Armadillo/beta-catenin repeat-containing protein
M136760	1.01	4.25	3	≥3rd	35.1.5	Pentatricopeptide repeat-containing protein, putative

* Genes not included in the major co-expression gene network but in the side minor networks

** The letter 'M' in gene IDs is abbreviated from 'MDP0000' in the original apple gene IDs used at the Genome Database for Rosaceae. Genes starting with 'G' are novel transcripts in the improved apple reference transcriptome (Bai et al. 2014)